Feature Correspondence via Graph Matching: Models and Global Optimization

Technical Report MSR-TR-2008-101

Lorenzo Torresani[1], Vladimir Kolmogorov[2], and Carsten Rother[1]

[1] Microsoft Research Ltd., Cambridge, UK, {ltorre, carrot}@microsoft.com
[2] University College London, UK, vnk@adastral.ucl.ac.uk

Abstract. In this paper we present a new approach for establishing correspondences between sparse image features related by an unknown non-rigid mapping and corrupted by clutter and occlusion, such as points extracted from a pair of images containing a human figure in distinct poses. We formulate this matching task as an energy minimization problem by defining a complex objective function of the appearance and the spatial arrangement of the features. Optimization of this energy is an instance of graph matching, which is in general a NP-hard problem. We describe a novel graph matching optimization technique, which we refer to as dual decomposition (DD), and demonstrate on a variety of examples that this method outperforms existing graph matching algorithms. In the majority of our examples DD is able to find the global minimum within a minute. The ability to globally optimize the objective allows us to accurately learn the parameters of our matching model from training examples. We show on several matching tasks that our learned model yields results superior to those of state-of-the-art methods.

1 Introduction

Feature correspondence is one of the fundamental problems of computer vision and is a key ingredient in a wide range of applications including object recognition, 3D reconstruction, mosaicing, motion segmentation, and image morphing. Several robust algorithms (see e.g. [1, 2]) exist for registration of images of static scenes and for visual correspondence under rigid motion. These methods typically exploit powerful constraints (e.g. epipolar constraints) to reduce the search space and disambiguate the correspondence problem. However, such constraints do not apply in the case of non-rigid motion or when matching different object instances. A popular approach in these cases is to discard the information about the spatial layout of features, and to find correspondences using appearance only. For example, many object recognition methods [3–9] represent images as orderless sets of local appearance descriptors, known as bags of features. Recent work [10] has suggested that for many correspondence problems, learned appearance-based models perform similarly or better than state-of-the-art structural models exploiting information about spatial arrangement of features. This is primarily due to the challenges posed by the optimization and training of structural models, which often require approximate solution of NP-hard problems. In this paper we contrast this theory, and demonstrate that a complex structural model for image matching

can be learned and optimized successfully. We cast the visual correspondence problem as an energy minimization task by defining a complex image matching objective depending on (i) feature appearance, (ii) geometric compatibility of correspondences, and (iii) spatial coherence of matched features. Additionally, we impose a uniqueness constraint allowing at most one match per feature. We introduce a novel algorithm to minimize this function based on the dual decomposition approach (DD) from combinatorial optimization, see e.g. [11–16]. The DD method works by maximizing a lower bound on the energy function. The value of the lower bound can be used to gauge the distance from the global minimum and to decide when to stop the optimization, in the event the global minimum cannot be found. For the majority of our examples DD finds the global minimum in reasonable time, and otherwise provides a solution whose cost is very close to the optimum. In contrast, previously proposed optimization methods such as [17, 18] often fail to compute good solutions for our energy function. Our experimental evaluation shows that the model and the algorithm presented in this paper can be applied to a wide range of image matching problems with results matching or exceeding those of existing algorithms [10, 19].

1.1 Relation to Previous Work

Models for feature matching Our approach is loosely related to algorithms that find visual correspondences by matching appearance descriptors under smooth, or piecewise smooth, spatial mappings. For example, Torr [20] describes a technique for estimating sparse correspondences using RANSAC under the assumption that the images contain a common set of rigidly moving objects. However, this piece-wise rigid motion assumption is not appropriate for deformable objects, such as human faces, or for different instances of an object class, such as different cars, related by highly non-linear mappings. Other approaches [21, 22] have proposed to constrain the correspondence problem by learning or hand-coding explicit models of how an object is allowed to deform using parametric 2D or 3D representations, such as linear eigenshapes or superquadrics. Unlike such approaches, our method does not make a parametric assumption about the transformation relating the input images, and thus can be used in a wider range of applications. Belongie et al. [19] inject spatial smoothness in the match by means of an iterative technique that alternates between finding correspondences using shape features, and computing a regularized transformation aligning the matching features. The shape descriptors are recomputed in each iteration after the warping. Since the objective is changed at each iteration, the convergence properties of this algorithm are not clear. Our approach is most closely related to the work of Berg et al. [23], and Leordeanu and Hebert [24], who formulate visual correspondence as a graph matching problem by defining an objective including terms based on appearance similarity as well as geometric compatibility between pairs of correspondences. Our model differs from those in [23, 24] in several ways. The methods proposed in [23] and [24] handle outliers by removing low-confidence correspondences from the obtained solutions. Instead, we include in our energy an explicit occlusion cost, as for example previously done in [25]. Thus our algorithm solves for the outliers as part of the optimization. We add to the objective a spatial coherence term, favoring spatial aggregation of matched features, which reduces the correspondence error on our examples. We also show that

geometric penalty functions defined in local neighborhoods provide more accurate correspondences than global geometric costs, such as those used in [23] and [24]. Finally, we use the method of Liu et al. [26] to learn the parameter values for the model from examples, thus avoiding the need of manual parameter tuning.

Graph matching optimization Graph matching is a challenging optimization problem which received considerable attention in the literature (see [27] for a comprehensive survey of methods). Proposed techniques include the graduated assignment algorithm of Gold and Rangarajan [28], spectral relaxation methods [24, 17], COMPOSE method of Duchi et al. [18]. Maciel and Costeira [29] reduce the problem to concave minimization and apply the exact method in [30]. Torr [25] and Schellewald and Schnörr [31] use semi-definite programming (SDP) relaxation for graph matching. Among these papers, only [29] and [31] report obtaining optimal (or near optimal) solutions. The method in [29] was tested only on a single example with quadratic costs. We conjecture that on practical challenging instances this method will suffer from an exponential explosion[3]. As shown in [17], the SDP relaxation approach in [31] scales quite poorly and is too expensive for problems of reasonable size.

2 Energy function

We now describe the energy function of our matching model. Let P' and P'' be the sets of features extracted from the two input images. We denote with $A \subseteq P' \times P''$ the set of potential assignments between features in the two sets. We will use the terms assignment and correspondence interchangeably to indicate elements of A. We represent a *matching configuration* between the two point sets as a binary valued vector $\boldsymbol{x} \in \{0,1\}^A$. Each correspondence $a \in A$ indexes an entry x_a in the vector $\boldsymbol{x}$. A correspondence a is active if $x_a = 1$, and it is inactive otherwise. We define an energy function $E(\boldsymbol{x})$ modeling our matching problem assumptions. This will allow us to formulate the matching task as minimization of $E(\boldsymbol{x})$. In this paper we consider matching problems where at most one active correspondence per feature is allowed. This requirement is known as the uniqueness constraint and it is commonly used in correspondence problems. In order to enforce this condition we define the constraint set M:

$$M = \{\boldsymbol{x} \in \{0,1\}^A \mid \sum_{a \in A(p)} x_a \leq 1 \quad \forall\, p \in P\} \tag{1}$$

where $P = P' \cup P''$ is the set of features from both images, and $A(p)$ is the set of correspondences involving feature p. The goal is to find the configuration $\boldsymbol{x} \in M$ minimizing $E(\boldsymbol{x})$. We define our energy as a weighted sum of four energy terms:

$$E(\boldsymbol{x}) = \lambda^{\text{app}} E^{\text{app}}(\boldsymbol{x}) + \lambda^{\text{occl}} E^{\text{occl}}(\boldsymbol{x}) + \lambda^{\text{geom}} E^{\text{geom}}(\boldsymbol{x}) + \lambda^{\text{coh}} E^{\text{coh}}(\boldsymbol{x}) \tag{2}$$

where $\lambda^{\text{app}}, \lambda^{\text{occl}}, \lambda^{\text{geom}}, \lambda^{\text{coh}}$ are scalar weights. We describe the energy terms below.

[3] The method in [29] first selects a *linear* function E^- which is an underestimator on the original objective function E, i.e. $E^-(\boldsymbol{x}) \leq E(\boldsymbol{x})$ for all feasible solutions $\boldsymbol{x}$. It then visits *all* feasible solutions $\boldsymbol{x}$ with $E^-(\boldsymbol{x}) \leq E(\boldsymbol{x}^*)$ where $E(\boldsymbol{x}^*)$ is the cost of the optimal solution. For each solution a linear program is solved.

Function $E^{\text{app}}(\boldsymbol{x})$ favors correspondences between features having similar appearance. We define this function as a sum of unary terms:

$$E^{\text{app}}(\boldsymbol{x}) = \sum_{a \in A} \theta_a^{\text{app}} x_a \ . \tag{3}$$

For an assignment $a = (p', p'') \in A$, θ_a^{app} is the distance between appearance descriptors (such as Shape Context [19]) computed at points p' and p'' in the respective images. We have used different features depending on the task at hand (see sec. 4).

The term $E^{\text{occl}}(\boldsymbol{x})$ imposes a penalty for unmatched features. We define $E^{\text{occl}}(\boldsymbol{x})$ to be the fraction of unmatched features in the smallest of the two feature sets. We can write this function as

$$E^{\text{occl}}(\boldsymbol{x}) = 1 - \frac{1}{\min\{|P'|, |P''|\}} \sum_{a \in A} x_a \tag{4}$$

by noting that $\sum_{a \in A} x_a$ is equal to the number of distinct matched features in P' and P'', $\forall \boldsymbol{x} \in M$. This result derives trivially from the uniqueness constraint.

The term $E^{\text{geom}}(\boldsymbol{x})$ is a measure of geometric compatibility between active correspondences. This term is similar to the distortion costs proposed in [23, 24]. Note, however, that the energy terms used in these previous approaches include distortion costs for all pairs of matched features, which results in energy functions penalizing any deviation from a global rigid transformation. Instead, our function $E^{\text{geom}}(\boldsymbol{x})$ measures geometric compatibility of correspondences only for *neighboring* features. We demonstrate that this model permits more flexible mappings between the two sets of features and yields more accurate correspondences. We use a "neighborhood system" N to specify the pairs of correspondences involved in our measure of geometric compatibility. N consists of all correspondence pairs defined over neighboring features:

$$N = \{\langle (p', p''), (q', q'') \rangle \in A \times A \mid p' \in N_{q'} \vee q' \in N_{p'} \vee p'' \in N_{q''} \vee q'' \in N_{p''}\} \tag{5}$$

where N_p indicates the set of K nearest neighbors of p (computed in the set of feature p), and K is a positive integer value controlling the size of the neighborhood, which we call *geometric neighborhood size*. $E^{\text{geom}}(\boldsymbol{x})$ is computed over pairs of active correspondences in the set N:

$$E^{\text{geom}}(\boldsymbol{x}) = \sum_{(a,b) \in N} \theta_{ab}^{\text{geom}} x_a x_b \tag{6}$$

where:

$$\theta_{ab}^{\text{geom}} = \eta(e^{\delta_{a,b}^2/\sigma_l^2} - 1) + (1 - \eta)(e^{\alpha_{a,b}^2/\sigma_\alpha^2} - 1) \tag{7}$$

$$\delta_{(p',p''),(q',q'')} = \frac{|||p' - q'|| - ||p'' - q''|||}{||p' - q'|| + ||p'' - q''||} \tag{8}$$

$$\alpha_{(p',p''),(q',q'')} = \arccos\left(\frac{p' - q'}{||p' - q'||} \cdot \frac{p'' - q''}{||p'' - q''||}\right) \tag{9}$$

Intuitively, $\theta^{\text{geom}}_{(p',p''),(q',q'')}$ computes the geometric agreement between neighboring correspondences (p', p''),(q', q'') by evaluating how well the segment $\overline{p'q'}$ matches the

segment $\overline{p''q''}$ in terms of both length and direction. The parameter η is a scalar value trading off the importance of preserving distances versus preserving directions.

The term $E^{\text{coh}}(\boldsymbol{x})$ favors spatial proximity of matched features. It incorporates our prior knowledge that matched features should form spatially coherent regions within each image, corresponding to common objects or parts in the image pair, in analogy to coherence on a pixel grid, used for example in image segmentation. We define the cost $E^{\text{coh}}(\boldsymbol{x})$ as the fraction of neighboring feature pairs with different occlusion status (this can be viewed as an MRF Potts model over feature occlusion). We now show how to write this function directly in terms of solution $\boldsymbol{x}$. Let N_P be the set of pairs of neighboring features in the two images:

$$N_P = \{(p,q) \in (P' \times P') \cup (P'' \times P'') \mid p \in N_q \vee q \in N_p\}. \tag{10}$$

Then we can express $E^{\text{coh}}(\boldsymbol{x})$ as a sum of unary and pairwise terms:

$$E^{\text{coh}}(\boldsymbol{x}) = \frac{1}{|N_P|} \sum_{(p,q)\in N_P} V_{p,q}(\boldsymbol{x}) \tag{11}$$

where:

$$V_{p,q}(\boldsymbol{x}) = \sum_{a\in A(p)} x_a + \sum_{b\in A(q)} x_b - 2 \sum_{a\in A(p), b\in A(q)} x_a x_b \; . \tag{12}$$

$V_{p,q}(\boldsymbol{x})$ is equal to 0 if p, q are either both matched or both unmatched; $V_{p,q}(\boldsymbol{x})$ is equal to 1 otherwise.

Feature correspondence as graph matching The problem defined above can be written as

$$\min_{\boldsymbol{x}\in M} E(\boldsymbol{x} \mid \bar{\theta}) = \sum_{a\in A} \bar{\theta}_a x_a + \sum_{(a,b)\in N} \bar{\theta}_{ab} x_a x_b \tag{13}$$

where the constraint set M is given by (1). This problem is often referred to as *graph matching* in the literature [28, 10]. Features P' and P'' are viewed as vertices of the two graphs. Pairwise term $\bar{\theta}_{ab} x_a x_b$ with $a = (p', p'')$, $b = (q', q'')$ encodes compatibility between edges (p', q'), (p'', q'') of the first and second graph, respectively, while unary term $\bar{\theta}_a x_a$ measures similarity between vertices p', p''.

We now address the question of how to optimize problem (13). Unfortunately, this problem is NP-hard [28]. We propose to use the *problem decomposition* approach (or *dual decomposition* - DD) for graph matching. Details are given in the next section.

3 Problem decomposition approach

On the high level, the idea is to decompose the original problem into several "easier" subproblems, for which we can compute efficiently a global minimum (or obtain a good lower bound). Combining the lower bounds for individual subproblems will then provide a lower bound for the original problem. The decomposition and the corresponding lower bound will depend on a parameter vector $\boldsymbol{\theta}$; we will then try to find a vector $\boldsymbol{\theta}$ that maximizes the bound. This approach is well-known in combinatorial optimization; sometimes it is referred to as "dual decomposition" [11]. It was applied to quadratic

pseudo-boolean functions (i.e. functions of binary variables with unary and pairwise terms) by Chardaire and Sutter [12]. Their work is perhaps the closest to the method in this paper. As in [12], we use "small" subproblems for which the global minimum can be computed exactly in reasonable time. Our choice of subproblems for graph matching, however, is different from [12]. In vision the decomposition approach is probably best known in the context of the MAP-MRF inference task. It was introduced by Wainwright et al. [13] who decomposed the problem into a convex combination of trees and proposed message passing techniques for optimizing vector $\boldsymbol{\theta}$. These techniques do not necessarily find the best lower bound (see [32] or review article [33]). Schlesinger and Giginyak [14, 15] and Komodakis et al. [16] proposed to use subgradient techniques [34, 11] for MRF optimization, which guarantee to converge to a vector $\boldsymbol{\theta}$ giving the best possible lower bound.

3.1 Graph matching via problem decomposition

We now apply this approach to the graph matching problem given by eq. (13). We decompose (13) into subproblems characterized by vectors θ^σ, $\sigma \in I$ with positive weights ρ_σ. (These weights are chosen a priori, and may affect the speed of convergence of the subgradient method in section 3.3.) Here I is a finite set of subproblem indexes. We will require the vector $\boldsymbol{\theta} = (\theta^\sigma \,|\, \sigma \in I)$ to be a *ρ-reparameterization* of the original parameter vector $\bar{\theta}$ [13], i.e.

$$\sum_{\sigma \in I} \rho_\sigma \theta^\sigma = \bar{\theta} \tag{14}$$

For each subproblem σ we will define a lower bound $\Phi_\sigma(\theta^\sigma)$ which satisfies

$$\Phi_\sigma(\theta^\sigma) \leq \min_{\boldsymbol{x} \in M} E(\boldsymbol{x} \mid \theta^\sigma) \tag{15}$$

It is easy to see that the function

$$\Phi(\boldsymbol{\theta}) = \sum_{\sigma \in I} \rho_\sigma \Phi_\sigma(\theta^\sigma) \tag{16}$$

is a lower bound on the original function. Indeed, if $\boldsymbol{x}^*$ is an optimal solution of (13) then from (14)-(16) we get

$$\Phi(\boldsymbol{\theta}) \leq \sum_{\sigma \in I} \rho_\sigma \min_{\boldsymbol{x} \in M} E(\boldsymbol{x} \mid \theta^\sigma) \leq \sum_{\sigma \in I} \rho_\sigma E(\boldsymbol{x}^* \mid \theta^\sigma) = E(\boldsymbol{x}^* \mid \bar{\theta})$$

In section 3.2 we will describe the subproblems that we use. For each subproblem σ we will do the following: (1) define constraints on vector θ^σ; (2) define the function $\Phi_\sigma(\theta^\sigma)$; (3) specify an algorithm for computing $\Phi_\sigma(\theta^\sigma)$. In section 3.3 we will discuss how to maximize the lower bound $\Phi(\boldsymbol{\theta})$ using the subproblem solutions. Finally, in section 3.4 we will describe how to obtain solution $\boldsymbol{x} \in M$ for our original problem.

3.2 Graph matching subproblems

Linear subproblem In our first subproblem, which we denote by the index "L", we require all pairwise terms to be zero: $\theta^L_{ab} = 0$ for $(a,b) \in N$. In such case problem (13) can be solved exactly in polynomial time, for example using the Hungarian algorithm [35]. (This is often known as the *linear assignment problem*.) We define $\Phi_L(\theta^L) = \min_{\boldsymbol{x} \in M} E(\boldsymbol{x} \mid \theta^L)$. To compute this minimum, we converted the problem to an instance of a minimum cost circulation with unit capacities and ran the successive shortest path algorithm [35]. This solves the problem using $O(|P| + |A|)$ Dijkstra shortest path computations in graphs with $|P| + 1$ nodes and $O(|P| + |A|)$ edges.

Maxflow subproblem In the second subproblem, which we denote by the index "M", we do not put any restrictions on the vector θ^M. To get a lower bound, we ignore the uniqueness constraint $\sum_{a \in A(p)} x_a \leq 1$ and leave only the discreteness constraint: $x_a \in \{0,1\}$. If the function $E(\boldsymbol{x} \mid \theta^M)$ is submodular (i.e. coefficients θ^M_{ab} are non-positive for all pairwise terms $(a,b) \in N$), then we can compute a global minimum using a maxflow algorithm. With arbitrary θ^M_{ab} the problem becomes NP-hard [36]. We use the *roof duality* relaxation [37] to get a lower bound $\Phi_M(\theta^M)$ on the problem. It can be defined as the optimal value of the following linear program:

$$\Phi_M(\theta^M) = \min \sum_{a \in A} \theta^M_a x_a + \sum_{(a,b) \in N} \theta^M_{ab} x_{ab} \tag{17}$$

$$\text{subject to } \begin{cases} 0 \leq x_a \leq 1 & \forall\, a \in A \\ x_{ab} \leq x_a, \quad x_{ab} \leq x_b, \quad x_{ab} \geq x_a + x_b - 1, \quad x_{ab} \geq 0 & \forall\, (a,b) \in N \end{cases}$$

This relaxation can be solved in polynomial time by computing a maximum flow in a graph with $2(|A| + 1)$ nodes and $O(|A| + |N|)$ edges [38, 36].

Local subproblems For our last set of subproblems we use an exhaustive search to compute the global minimum (see Appendix A for details). Thus, we need to make sure that subproblems are sufficiently small. We use the following technique. For each point $p \in P$ we choose $N^d_p \subseteq P$ to be the set of K^d nearest points in the same image where K^d is a small constant, e.g. 2 or 3. (The superscript d stands for "decomposition".) We then consider the subproblem which involves only assignments in the set $A(N^d_p) = \{(p', p'') \in A \mid p' \in N^d_p \vee p'' \in N^d_p\}$ and the edges between those assignments. More precisely, we require vector θ^p corresponding to this subproblem to satisfy the following constraints:

$$\begin{aligned} \theta^p_a &= 0 \;\text{ if }\; a \notin A(N^d_p), \\ \theta^p_{ab} &= 0 \;\text{ if }\; a \notin A(N^d_p) \;\text{ or }\; b \notin A(N^d_p). \end{aligned}$$

These constraints imply that we can fix assignments $a \in A - A(N^d_p)$ to 0 when computing the minimum $\min_{\boldsymbol{x} \in M} E(\boldsymbol{x} \mid \theta^p)$. Then we get a graph matching problem where the set of points in one of the images is N^d_p.

3.3 Lower bound optimization

In the previous section we described constraints on vector $\boldsymbol{\theta}$ and a lower bound $\Phi(\boldsymbol{\theta})$ consisting of $|P|+2$ subproblems. It can be seen that Φ is a concave function of $\boldsymbol{\theta}$. Furthermore, the constraints on $\boldsymbol{\theta}$ yield a convex set Ω. This set is defined by the reparameterization equation (14) and constraints on individual subproblems $\theta^\sigma \in \Omega_\sigma$ given by equalities $\theta^\sigma_a = 0$, $\theta^\sigma_{ab} = 0$ for certain assignments a and edges (a,b). Let $I_a, I_{ab} \subseteq I$ be the subsets of subproblem indexes for which elements θ^σ_a, θ^σ_{ab}, respectively, are **not** constrained to be 0. Thus, assignment $a \in A$ is involved in subproblems $\sigma \in I_a$, and edge $(a,b) \in N$ is involved in subproblems $\sigma \in I_{ab}$.

Similar to [12, 14–16], we used a projected subgradient method [34, 11] for maximizing $\Phi(\boldsymbol{\theta})$ over Ω. One iteration is given by

$$\theta := \mathcal{P}_\Omega(\boldsymbol{\theta} + \lambda \mathbf{g})$$

where $\mathcal{P}_\Omega$ is the operator that projects a vector to Ω, $\mathbf{g}$ is a subgradient of $\Phi(\boldsymbol{\theta})$ and $\lambda > 0$ is a step size.

Projection To project vector $\boldsymbol{\theta}$ to Ω, we first compute vector $\hat{\theta} = \sum_\sigma \rho_\sigma \theta^\sigma$ and then update $\boldsymbol{\theta}$ as follows: $\theta^\sigma_a := 0$ for $\sigma \in I - I_a$, $\theta^\sigma_{ab} := 0$ for $\sigma \in I - I_{ab}$,

$$\theta^\sigma_a := \theta^\sigma_a + \rho_\sigma \frac{\bar{\theta}_a - \hat{\theta}_a}{\sum_{\sigma \in I_a} \rho^2_\sigma} \qquad \forall\, \sigma \in I_a,$$

$$\theta^\sigma_{ab} := \theta^\sigma_{ab} + \rho_\sigma \frac{\bar{\theta}_{ab} - \hat{\theta}_{ab}}{\sum_{\sigma \in I_{ab}} \rho^2_\sigma} \qquad \forall\, \sigma \in I_{ab}.$$

Subgradient computation A subgradient of function $\Phi(\boldsymbol{\theta})$ is given by

$$\mathbf{g} = \sum_{\sigma \in I} \rho_\sigma \mathbf{g}^\sigma$$

where $\mathbf{g}^\sigma$ is a subgradient of function $\Phi_\sigma(\theta^\sigma)$. If the latter function is the global minimum of $E(\boldsymbol{x} \mid \theta^\sigma)$ (which is the case for $\sigma \in I - \{M\}$) then we can take $g^\sigma_a = x^\sigma_a$, $g^\sigma_{ab} = x^\sigma_a x^\sigma_b$ where $\boldsymbol{x}^\sigma$ is a global minimizer of $E(\boldsymbol{x} \mid \theta^\sigma)$. For the maxflow subproblem a subgradient can be computed as $\mathbf{g}^M = \boldsymbol{x}^M$ where $\boldsymbol{x}^M$ is an optimal solution of linear program (17). The method in [38] produces a half-integer optimal solution where $x^M_a \in \{0, 0.5, 1\}$ for all assignments a and x^M_{ab} is determined as follows: if $(x^M_a, x^M_b) \neq (0.5, 0.5)$ then $x^M_{ab} = x^M_a x^M_b$, otherwise $x^M_{ab} = 0$ if $\theta^M_{ab} \leq 0$ (i.e. the corresponding term is submodular) and $x^M_{ab} = 0.5$ if $\theta^M_{ab} > 0$.

Step size An important issue in the subgradient method is the choice of the step size λ. We used an adaptive technique described in [39, 11]. We set $\lambda = \alpha(\Phi(\boldsymbol{\theta}^*) + \delta - \Phi(\boldsymbol{\theta}))/||\mathbf{g}||^2$ where α is a constant (1 in our experiments), $\boldsymbol{\theta}^*$ is the best vector found so far (i.e. the vector giving the best lower bound), and δ is a positive number which is updated as follows: if the last iteration improved the best lower bound $\Phi(\boldsymbol{\theta}^*)$ then δ is increased by a certain factor (1.5 in our experiments), otherwise it is decreased by a certain factor (0.95).

Restarting the subgradient method In our implementation we also used the following technique borrowed from [40]. If the best value of the lower bound $\Phi(\boldsymbol{\theta}^*)$ has not

changed during γ iterations then we replace $\boldsymbol{\theta}$ with $\boldsymbol{\theta}^*$. In the beginning $\gamma = 20$, and after every restart it is updated as $\gamma := \min\{\gamma + 10, 50\}$.

3.4 Solution computation

To conclude the description of the method, we need to specify how to obtain solution $\boldsymbol{x} \in M$. If the linear subproblem is included in the decomposition then it is natural to use its minimum $\boldsymbol{x}^L$ in each iteration, since $\boldsymbol{x}^L$ is guaranteed to satisfy the uniqueness constraint. However, we excluded this subproblem for the experiments presented in this paper (for reasons explained below). We computed the solution in a given iteration as follows: starting with labeling $\boldsymbol{x} = 0$, we go through local subproblems $\sigma \in I - \{L, M\}$ and assignments a involved in σ (in a fixed order), set $x_a = 1$ if $x^\sigma_a = 1$ and this operation preserves the uniqueness constraint on $\boldsymbol{x}$.

We maintain the solution with the smallest energy computed so far, and output it as a result of the method.

3.5 Properties of decomposition

Of course, it is not necessary to use all subproblems described in section 3.2. The only requirement is that each assignment $a \in A$ and edge $(a, b) \in N$ should be covered by at least one subproblem (i.e. I_a and I_{ab} should be non-empty), otherwise the projection operation would be undefined. In this section we study how the choice of subproblems affects the optimal value of the lower bound $\max_{\boldsymbol{\theta} \in \Omega} \Phi(\boldsymbol{\theta})$. Without loss of generality we can assume $\rho_\sigma = 1$ for $\sigma \in I$. (Indeed, weights ρ_σ may affect the speed of the subgradient method, but they do not affect the value of the optimal bound since the transformation $\rho_\sigma := \rho_\sigma / \gamma$, $\theta^\sigma := \theta^\sigma \cdot \gamma$ with $\gamma > 0$ preserves the bound.)

First, we compare the bound provided by the decomposition method with the following technique which we call *QPBO*:

1. For each constraint $\sum_{a \in A(p)} x_a \leq 1$ of the set M add pairwise terms $C x_a x_b$ for all pairs of assignments $a, b \in A(p)$, $a \neq b$ where C is a large constant ensuring that $x_a x_b = 0$ in the optimal solution. Let $E'(\boldsymbol{x})$ be the function that we obtain. Clearly, the minimization problem $\min_{\boldsymbol{x} \in \{0,1\}^A} E'(\boldsymbol{x})$ is equivalent to (13).
2. Minimizing E' is an instance of a *quadratic pseudo-boolean optimization* problem [36]. Apply the roof duality relaxation [37, 36] to get a lower bound on (13).

Lemma 1. *If the set I includes the linear and maxflow subproblems then the optimal value of the lower bound $\Phi(\boldsymbol{\theta}^*)$ is the same as or larger than the QPBO bound.*

A proof is given in Appendix B. In this proof we derive the LP relaxation solved by the decomposition approach in the case when $I = \{L, M\}$.

The next lemma shows that the linear and maxflow subproblems are often not essential. (A proof is given in Appendix C.)

Lemma 2. *(a) Suppose that for each point $p \in P$ there exists a local subproblem $\sigma \in I - \{L, M\}$ which covers all assignments in $A(p)$, i.e. $\sigma \in I_a$ for all $a \in A(p)$. Then adding or removing the linear subproblem will not affect the optimal value of the lower bound of the decomposition approach.*

(b) Suppose that each assignment $a \in A$ and each edge $(a, b) \in N$ are covered by at least one local subproblem, i.e. there exist subproblems $\sigma \in I - \{L, M\}$ with $\sigma \in I_a$ and subproblems $\sigma \in I - \{L, M\}$ with $\sigma \in I_{ab}$. Then adding or removing the maxflow subproblem will not affect the optimal value of the lower bound.

It can be seen that our choice of local subproblems always satisfies conditions of part (a). Thus, the linear subproblem would not help (assuming that we can compute the optimal lower bound). We described this subproblem partly because it was used in previous work: in [18] the authors computed *exact* min-marginals for the linear subproblem in the belief propagation framework.

As for part (b), the answer depends on the structures of the neighborhood systems N_p used for constructing the energy function and N_p^d used for constructing local subproblems. Recall that N_p and N_p^d are controlled by parameters K and K^d, respectively. If $K \leq K^d$ then conditions of part (b) are always satisfied, otherwise some edges may not be covered, and so including the maxflow subproblem may improve the optimal bound.

4 Experimental results

In most of our experiments we learned problem-specific parameters of our energy model from ground truth correspondences. The learning procedure was initialized using default parameters corresponding to uniform values for the weights λ_i, and variance values $\sigma_l^2 = 0.5$, $\sigma_\alpha^2 = 0.9$. We now describe the learning technique.

4.1 Model learning

The energy model defined in Equation (2) is parameterized by a set of parameters, denoted here with $\boldsymbol{\psi} = \{\lambda^{\text{app}}, \lambda^{\text{occl}}, \lambda^{\text{geom}}, \lambda^{\text{coh}}, \eta, \sigma_l^2, \sigma_\alpha^2\}$. In addition, the energy depends on input features sets P' and P'' extracted from the images. Here we highlight this dependence on parameters and input points, by writing the energy function as $E(\boldsymbol{x}; P', P'', \boldsymbol{\psi})$. We now consider the problem of learning parameters $\boldsymbol{\psi}$ from a set of n training matching examples defined by pairs of feature sets $\{(P'_1, P''_1), ..., (P'_n, P''_n)\}$ and "ground truth" correspondences $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$, typically specified by the user. We use the Nonlinear Inverse Optimization (NIO) algorithm described in [26]. The objective of this method corresponds in our case to minimizing the gap in energy value between the user-provided ground truth correspondences and the matching configurations estimated via energy optimization. In other words, we minimize the following objective $G(\boldsymbol{\psi})$:

$$G(\boldsymbol{\psi}) = \sum_{i=1}^{n} E(\boldsymbol{x}_i; P'_i, P''_i, \boldsymbol{\psi}) - \min_{\boldsymbol{x} \in M} E(\boldsymbol{x}; P'_i, P''_i, \boldsymbol{\psi}).$$

Let $\boldsymbol{\lambda} = \{\lambda^{\text{app}}, \lambda^{\text{occl}}, \lambda^{\text{geom}}, \lambda^{\text{coh}}\}$, and $\boldsymbol{\sigma}^2 = \{\sigma_l^2, \sigma_\alpha^2\}$. In order to avoid degenerate solutions with $\boldsymbol{\lambda} = \mathbf{0}$, and to obtain positive values for the term weights λ_i and variances σ_i^2, we define reparameterizations $\lambda_i = e^{\nu_i} / \sum_j e^{\nu_j}$, $\sigma_i^2 = e^{\epsilon_i}$, and minimize G with

respect to $\{\boldsymbol{\epsilon}, \boldsymbol{\nu}\}$ instead of $\boldsymbol{\psi}$. As in [26], we optimize G via gradient descent with line search. The gradient of G is locally approximated by optimizing the energies given the current estimate of the parameters in each iteration.

4.2 Algorithms

In our experiments we compare the following algorithms:

DD We used $K^d = \min\{K, 4\}$, where K is the geometric neighborhood size. Motivated by results in sec. 3.5, we did not use the linear subproblem. We set $\rho_\sigma = 1$ for all other subproblems σ. We used a maximum of 10000 iterations, and stopped earlier if the gap between the lower bound and the cost became smaller than 10^{-6}.

FUSION This technique was introduced in [41] for MRF optimization with multiple labels. We propose to use it for graph matching as follows. First, we generate 256 solutions by applying one pass of coordinate descent (ICM) to zero labeling using random orders. (Different orders of visiting assignments usually yield different solutions.) We then "fuse" together pairs of solutions using the binary tree structure until a single solution remains. Fusion of solutions $\boldsymbol{x}'$, $\boldsymbol{x}''$ is defined as follows. First, we fix all assignments $a \in A$ for which $\boldsymbol{x}'$ and $\boldsymbol{x}''$ agree, i.e. $x'_a = x''_a$. Then we convert the obtained graph matching problem to a quadratic pseudo-boolean optimization problem as described in section 3.5. Finally, we run the QPBO-PI method [42] starting either with labeling $\boldsymbol{x}'$ if $E(\boldsymbol{x}' \mid \bar{\theta}) < E(\boldsymbol{x}'' \mid \bar{\theta})$ or with $\boldsymbol{x}''$ otherwise. The produced solution $\boldsymbol{x}$ is guaranteed to have the same or smaller cost than the costs of $\boldsymbol{x}'$ and $\boldsymbol{x}''$.

Below we show plots of the energy as a function of time. Clearly, these plots depend on the order of fusions. We used the following order: we always pick the leftmost node of the binary tree whose parents are available for fusion. Thus, the plots are independent of the number of initial solutions (256 in our case).

BP We converted graph matching to a quadratic pseudoboolean optimization problem and ran max-product belief propagation algorithm[4]. We also tested applying the roof duality approach instead of BP, but results were quite discouraging (see below).

SMAC We ran the spectral relaxation method of Cour et al. [17], using the graduated assignment algorithm [28] for discretization. Since SMAC imposes affine constraints on the solution, we applied this algorithm only to datasets without outliers, where the one-to-one affine constraint is satisfied. In principle, SMAC could handle outliers by the introduction of dummy nodes. However, this would increase the number of variables and potentially make the problem harder to solve.

COMPOSE We reimplemented the algorithm in [18]. The problem was cast as assigning a label from the set $A(p') \cup \{\text{“}occlusion''\}$ to each point $p' \in P'$. Min-marginals for the linear subnetwork were computed via $O(|A| + |P'|)$ calls to the Dijkstra algorithm. As in [18], we used Residual Belief Propagation (RPB) [43] with damping=0.3 for computing pseudo min-marginals for the "smoothness" subnetwork containing pairwise terms $\theta_{ab} x_a x_b$. However, in our experiments messages did not converge, so we set an additional termination criterion for RBP: we stop it after passing $20|N|$ messages. As in [18], we computed the configuration by looking at individual messages at each

[4] We used the code from http://www.adastral.ucl.ac.uk/~ vladkolm/papers/TRW-S.html

node. We did not use damping for the outer loop since otherwise the produced configurations usually did not satisfy the uniqueness constraint. We also tested informally the COMPOSE method with our representation which labels each assignment as 0 or 1. To compute min-marginals for the smoothness subnetwork we tried both a maxflow algorithm (in the case of submodular potentials) and RBP. However, it did not seem to improve the results, and the issue with convergence remained.

HUNG As in [10], we also tested the Hungarian algorithm using an energy consisting only of linear terms. On problems with occlusions, we used our occlusion cost in addition to the appearance energy term, i.e. $E^{\text{HUNG}}(\boldsymbol{x}) = \lambda^{\text{app}} E^{\text{app}}(\boldsymbol{x}) + \lambda^{\text{occl}} E^{\text{occl}}(\boldsymbol{x})$.

4.3 Comparative results

Hotel sequence: wide baseline matching. In this subsection we report results on the CMU 'hotel' sequence [5]. As in previous work [10], we use this dataset to assess the performance of graph matching methods, and ignore the rigid motion constraint that could be exploited using alternative wide-baseline matching algorithms [20]. We reproduce the experimental setup described in [10] using the same manual labeling of 30 landmark points, and the same subset of 105 frame pairs. As in this previous work, we adopt as unary terms the distances between Shape Context descriptors. However, we replace the pairwise terms proposed in [10] with our geometric energy function $E^{\text{geom}}(\boldsymbol{x})$, using $K = 2$. Due to the absence of outliers, we remove $E^{\text{coh}}(\boldsymbol{x})$ from our energy and use a large constant value for λ^{occl}. We set the remaining parameters to default values, as defined above. We set $A = P' \times P''$. Figure 1(a) shows the matching error obtained by optimizing this model with different methods. We include in the plot also the performance of HUNG. Note the very large variance in matching performance, with BP and DD being the best methods with errors approaching 0%. Note that the error obtained with our model and our optimization is over 50 times smaller than the errors recently reported in [10]. On this dataset DD found always the global minimum and in each case within a minute (see Figure 1(b)). We found that on this sequence QPBO does not provide any labeling at all. Figure 1(c) illustrates performance versus runtime on one image pair (frame 1 and 64). In this plot we indicate convergence to a global minimum with a green circle. BP does well on this sequence, nearly matching the minimization performance of DD, at a reduced cost.

Matching MNIST digits. Here we describe experiments on images of handwritten digits from the MNIST dataset [44]. For training, we randomly sampled from this dataset one image pair for each of the 10 digits. We repeated the same procedure to generate a test set of 10 pairs of same digits. From each pair we extracted point sets P' and P'' by uniformly sampling 100 points along the Canny edges of each image, using the procedure described in [19]. We defined the unary potentials $\theta^{\text{app}}_{(p',p'')}$ to be the Euclidean distances between Shape Context descriptors computed at points p', p''. We formed the set of candidate assignments $A \in P' \times P''$ by selecting the 5 most similar features, in terms of Shape Context distance, for each point $p \in P$. We collected ground truth correspondences in the set $(P' \times P'')$ for each of the 20 image pairs. The parameters of our model were learned from the 10 training image pairs with

[5] Available at: http://vasc.ri.cmu.edu/idb/html/motion/hotel/index.html.

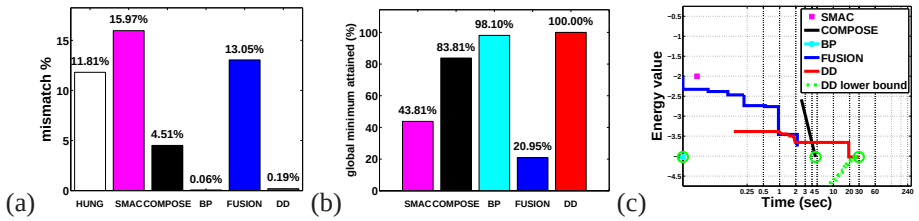

Fig. 1. Results on the Hotel sequence ($|P'| = |P''| = 30, |A| = 900$). (a) Mismatch percentages of HUNG and different optimizations applied to our energy model. (b) Frequency of convergence to global minimum. (c) Energy minimization versus time.

NIO. Figure 2(a) shows that the matching accuracy on the test set critically depends on the ability to globally optimize the energies during the model learning stage. The left plot reports the frequency of convergence to a global minimum during learning, plotted as a function of K, the geometric neighborhood size. The second plot shows the test set matching error of DD with learned versus default parameters. Matching error here is measured as percentage of incorrect correspondences with respect to ground truth[6]. We can see that the matching is much more accurate when using the parameters for which DD reached more frequently global optimality during learning. Interestingly, although the frequency of global minimum convergence increases slightly when varying K from 2 to 4, the matching error remains roughly the same. This suggests that geometric penalty terms defined over small neighborhoods are sufficient to spatially regularize the correspondences. Thus, models involving geometric costs defined over all pairs of matched features, such as those used in [23, 24], may be unnecessarily restrictive for many applications, in addition to being more difficult to optimize.

Given these results, we have used the model learned with $K = 3$ for the MNIST experiments described below. Figure 2(b) shows the normalized energy values obtained by different optimization methods on the test set. For each family of results we performed an *additive* normalization so that for each image pair the energy of the best method becomes a fixed number. On 9 out of the 10 test image pairs, DD reaches global optimality, and provides the minimum energy value on all examples. FUSION, BP, and COMPOSE find the global minimum only on 2 images. FUSION finds solutions with energy values very close to those obtained by DD. COMPOSE and BP provide considerably higher energy values on some of the examples.

We have also attempted to minimize the energy by running the QPBO algorithm [37, 38, 36, 42] on the equivalent quadratic pseudoboolean optimization problem described in section 3.5. This algorithm produces partial labelings that are part of a global optimum. However, we found that on our MNIST matching problems, QPBO labeled on average only 16% of the correspondences, with only 0.12% of these assignments cor-

[6] In order to account for a certain degree of inherent ambiguity in the selection of ground truth correspondences on these images, we did not consider it an error if a point was assigned to any of the 3-nearest neighbors of the correct feature. Declaring a point with a ground truth correspondence an outlier (or viceversa) was counted as error.

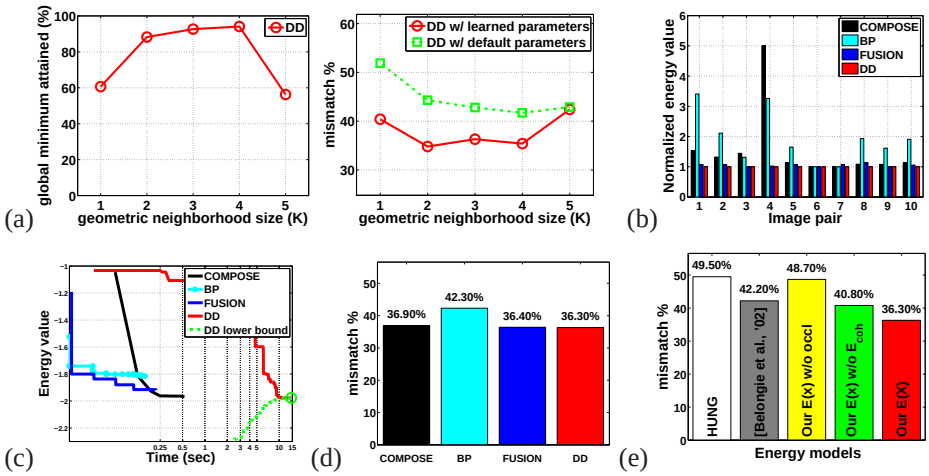

Fig. 2. Experimental results on MNIST digits ($|P'| = |P''| = 100, |A| = 695$, on average). (a) Correlation between learning accuracy and matching performance: the left plot shows the frequency of global minimum convergence during learning versus K; the right plot shows mismatch error on test set. (b) Normalized energy values. (c) Optimization performance versus runtime. (d) Mismatch error comparison between different optimization methods using our energy model. (e) Mismatch error using different energy models.

responding to active correspondences. We also tried to apply the PROBE method [45, 42] to get more labeled nodes. However, in practice we were unable to do so, due to the high computational cost of running PROBE on our problem instances.

Figure 2(c) shows minimization performance as a function of time, evaluated on a sample image pair. Figure 2(d) shows the correspondence accuracy obtained by optimizing our energy with the different methods. Again, we find that DD and FUSION yield the best accuracy. The parameters used for the energy of HUNG were learned from the training examples using NIO with Hungarian matching for optimization. We also evaluated variations of the energy model defined in Equation (2) obtained by using only the linear tearms (HUNG), by dropping the spatial coherence term, and by forcing all points to be matched (implemented by fixing λ^{occl} to a large value). The parameters of these modified models were learned again with NIO, using DD for both training and testing. We see from Figure 2(e) that both the spatial coherence prior, as well as the occlusion cost, improve the matching accuracy. On these instances the simple appearance-based model used by HUNG gives poor accuracy. We also report the matching error given by the model and optimization method of Belongie et al., which was applied to MNIST digit examples in [19]. For this experiment, we use the source code provided by the authors and the settings described in [19]. Our approach performs better than this state-of-the-art method.

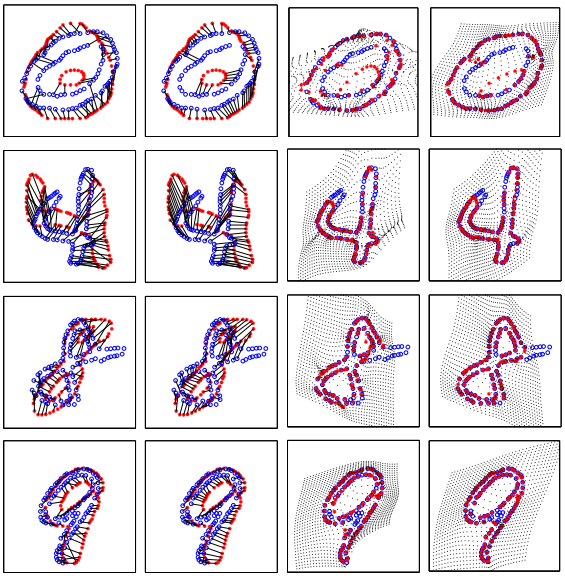

Fig. 3. Correspondences obtained with (a) Hungarian matching and (b) DD. The last two columns show grids warped by the thin-plate spline transformations computed from the correspondences of (c) Hungarian matching and (d) DD.

Figure 3 illustrates some of the matches obtained with DD and HUNG. Our model yields more accurate and geometrically consistent correspondences.

Estimating long range non-rigid motion. In this subsection we describe results on the task of estimating large-disparity motion. For this experiment we used four (time-separated) video frames of a child jumping. We matched each image to every other image, for a total of six matches. The motion between any pair of these pictures is very large and highly non-rigid. There is self-occlusion created by the motion of arms and torso, and occlusion due to a tricycle positioned between the child and the camera. Feature points were extracted by running the Harris corner detector on each image. We used Euclidean distances of geometric blur descriptors [23] computed at each feature point, both for selecting assignments in A (by choosing the five most similar features for each point $p \in P$) as well as for calculating the unary terms of our energy. We learned the parameters in our model by applying the NIO algorithm to ground truth correspondences of two image pairs from a separate sequence containing the same child walking. Here we report results using $K = 6$. Figure 4 shows two matching examples from this experiment and correspondences found with HUNG and DD. Note the ability of our system to cope well with occlusion and multiple motions. DD converged to a global minimum on all the image pairs in this experiment (see Figure 5(a,b)). Figure 5(c) reports the correspondence errors (including mismatches as well as missed assignments).

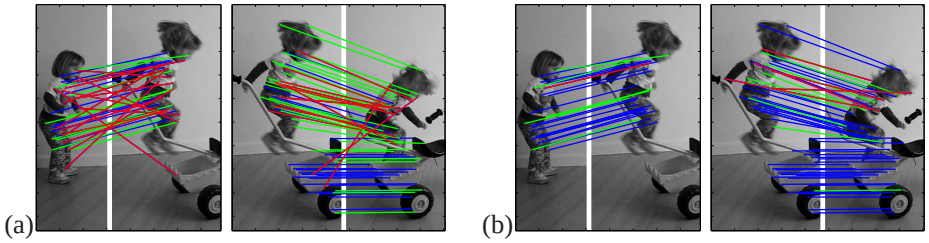

Fig. 4. Estimating human motion ($|P'| = 118, |P''| = 172, |A| = 1128$ on average). Correspondences computed with (a) the Hungarian method and (b) DD. Correct correspondences are shown in blue, missed assignments in green, and mismatches in red.

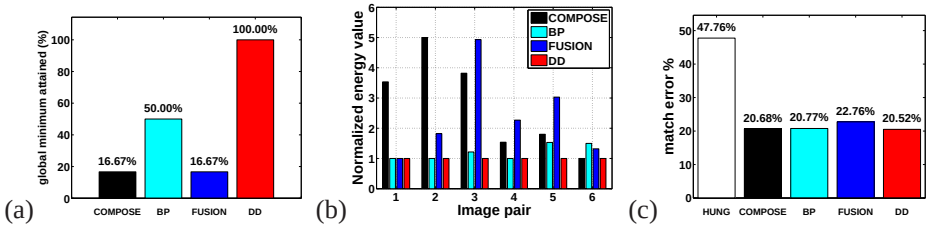

Fig. 5. Experiments on human motion frames. (a) Frequency of convergence to global minimum. (b) Normalized energy values. (c) Correspondence error.

Matching faces. We also carried out experiments on a set of 8 face images of distinct individuals with different facial expressions. We used 2 of these images for learning the model parameters given their manually labeled correspondences. We then exhaustively matched the remaining 6 images, for a total of 15 test image pairs. Point sets P' and P'' and candidate assignments A for each image pair were formed by matching geometric blur descriptors [23] computed along Canny edges in each image, using an iterative procedure. Starting from empty sets $P' = P'' = A = \{\}$, we alternate selection of a new point p from either the left or the right image, by choosing the edge point (among those not yet considered) having minimum geometric blur distance to points in the other image. We add the 3 best assignments involving point p to A, and the corresponding points to P' and P''. We then introduce an inhibition window around point p so that no other points in that neighborhood will be selected. We repeat this procedure 600 times. On average this yields point sets with more than 900 points in each image, and a set A with over 1700 potential assignments. Here we used geometric neighborhood size $K = 6$, and defined again the unary term to be the Euclidean distance between geometric blur descriptors. Figure 6(a) shows the normalized energy values obtained with the optimization methods in our comparison. FUSION is the best performing method 93.33% of the times, while DD is the best on the remaining 6.67% cases. On all image pairs DD and FUSION obtain very similar values, but on none of these challenging matches they are able to reach the lower bound. By dropping the value of K to 4, we

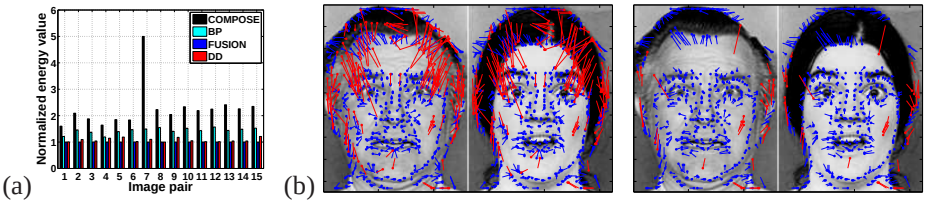

Fig. 6. Results on face images ($|P'| = 922, |P''| = 915, |A| = 1782$, on average). (a) Normalized energy values. (b) Correspondences found with Hungarian matching (left) and DD (right).

found that DD can reach the global minimum in most of the cases, although the correspondences are slightly less accurate than when using $K = 6$. Here BP performs rather poorly. Figure 6(b) illustrates the correspondences found by HUNG and DD on one of these image pairs. Assignments are shown as feature displacements in each image. Red lines denote incorrect assignments.

5 Conclusions

We have introduced novel models and optimization algorithms for feature correspondence. We believe to be the first to demonstrate graph matching techniques capable of reaching global optimality on various real-world image matching problems. As a future work, we plan to replace exhaustive search for local subproblems with a branch-and-bound method, as in [12]. We hope that this may speed up substantially the DD method.

Appendix A: exhaustive search

We computed the global minimum for local subproblems used in the decomposition approach as follows. Assume that the set P' has a smaller size than P'' (the other case is symmetric). First, we select on ordering of points in P'. We then use a depth-first search to go through all labelings $\boldsymbol{x} \in M$. We start with the zero labeling in which all assignments are passive. At depth d ($d = 1, \ldots, |P'|$) we pick d-th point p in P' and explore $|A(p)| + 1$ possible branches for p. (In each branch we either make one of assignments in $A(p)$ active, or declare all assignments in $A(p)$ to be passive.) If we detect a violation of the uniqueness constraint in P'' then we backtrack. For each depth d we maintain the cost of the current labeling. Updating this cost at depth d takes $O(d)$ time. (For this we need to store an $|A| \times |A|$ matrix of costs in the memory.)

Appendix B: proof of lemma 1

Let us write down the linear program corresponding to the QPBO method. The roof duality relaxation for function E is given by equation (17). Adding pairwise terms $Cx_a x_b$

for $a, b \in A(p)$, $a \neq b$ to function (13) will affect the relaxation (17) as follows: linear terms Cx_{ab} will be added to function (17), and corresponding constraints will be imposed (see (17)). Since C is a large constant, new variables x_{ab} will be forced to 0. Therefore, we arrive at the following linear program:

$$\min \quad \sum_{a \in A} \bar{\theta}_a x_a + \sum_{(a,b) \in N} \bar{\theta}_{ab} x_{ab} \tag{18}$$
$$\text{subject to} \begin{cases} x_a + x_b \leq 1 & \forall\, a, b \in N(p), p \in P, a \neq b \\ 0 \leq x_a \leq 1 & \forall\, a \in A \\ x_{ab} \leq x_a, x_{ab} \leq x_b & \forall\, (a,b) \in N \\ x_{ab} \geq x_a + x_b - 1 & \forall\, (a,b) \in N \\ x_{ab} \geq 0 & \forall\, (a,b) \in N \end{cases}$$

Let us now derive the relaxation solved by the decomposition approach with the linear and maxflow subproblems, i.e. with $I = \{L, M\}$.It is well-known [35] that the optimal value of the linear matching problem $\Phi_L(\theta^L)$ is equal to the optimal value of the following linear program:

$$\begin{aligned} & \min \quad \sum_{a \in A} \theta^L_a x_a \\ & \text{subject to} \begin{cases} \sum_{a \in A(p)} x_a \leq 1 & \forall\, p \in P \\ x_a \geq 0 & \forall\, a \in A \end{cases} \\ = & \max \quad \sum_{p \in P} -\mu_p \\ & \text{subject to} \begin{cases} -\mu_p - \mu_q \leq \theta^L_a & \forall\, a = (p,q) \in A \\ \mu_p \geq 0 & \forall\, p \in P \end{cases} \end{aligned} \tag{19}$$

Similarly, the lower bound for the maxflow subproblem $\Phi_M(\theta^M)$ can be written as the dual problem to (17):

$$\max \quad \sum_{a \in A} -\lambda_a + \sum_{(a,b) \in N} -\lambda_{ab} \tag{20}$$
$$\text{subject to} \begin{cases} -\lambda_a + \sum_{(a,b) \in N} \left[\bar{\lambda}_{ab} - \lambda_{ab}\right] \leq \theta^M_a & \forall\, a \in A \\ -\bar{\lambda}_{ab} - \bar{\lambda}_{ba} + \lambda_{ab} \leq \theta^M_{ab} & \forall\, (a,b) \in N \\ \lambda_a \geq 0 & \forall\, a \in A \\ \bar{\lambda}_{ab} \geq 0, \bar{\lambda}_{ba} \geq 0, \lambda_{ab} \geq 0 & \forall\, (a,b) \in N \end{cases}$$

Here we denoted $\bar{\lambda}_{ab}$ and λ_{ab} to be the dual variables for the constraints $x_{ab} \leq x_a$ and $x_{ab} \geq x_a + x_b - 1$, respectively. Note that $\bar{\lambda}_{ab}$ and $\bar{\lambda}_{ba}$ are distinct variables, while λ_{ab} and λ_{ba} denote the same variable.

Using (19) and (20), we can write the optimal lower bound of the decomposition approach as follows:

$$
\begin{aligned}
&\max_{\substack{\theta^L \in \Omega^L \\ \theta^L + \theta^M = \bar{\theta}}} \Phi_L(\theta^L) + \Phi_M(\theta^M) \\
&= \max_{\theta : \theta_{ab} = 0} \Phi_L(-\theta) + \Phi_M(\bar{\theta} + \theta) \\
&= \max_{\theta, \mu, \lambda, \bar{\lambda}} \quad \sum_{p \in P} -\mu_p + \sum_{a \in A} -\lambda_a + \sum_{(a,b) \in N} -\lambda_{ab} \\
&\qquad \text{subject to} \begin{cases}
-\mu_p - \mu_q \leq -\theta_a & \forall\, a = (p,q) \in A \\
-\lambda_a + \sum\limits_{(a,b) \in N} \left[\bar{\lambda}_{ab} - \lambda_{ab} \right] \leq \bar{\theta}_a + \theta_a & \forall\, a \in A \\
-\bar{\lambda}_{ab} - \bar{\lambda}_{ba} + \lambda_{ab} \leq \bar{\theta}_{ab} & \forall\, (a,b) \in N \\
\mu_p \geq 0 & \forall\, p \in P \\
\lambda_a \geq 0 & \forall\, a \in A \\
\bar{\lambda}_{ab} \geq 0, \bar{\lambda}_{ba} \geq 0, \lambda_{ab} \geq 0 & \forall\, (a,b) \in N
\end{cases}
\end{aligned}
$$

We can eliminate θ_a from the first and the second constraint and combine them into one constraint, then we obtain

$$
\begin{aligned}
&\max_{\mu, \lambda, \bar{\lambda}} \quad \sum_{p \in P} -\mu_p + \sum_{a \in A} -\lambda_a + \sum_{(a,b) \in N} -\lambda_{ab} \\
&\qquad \text{subject to} \begin{cases}
-\mu_p - \mu_q - \lambda_a + \sum\limits_{(a,b) \in N} \left[\bar{\lambda}_{ab} - \lambda_{ab} \right] \leq \bar{\theta}_a & \\
\qquad\qquad \forall\, a = (p,q) \in A & \\
-\bar{\lambda}_{ab} - \bar{\lambda}_{ba} + \lambda_{ab} \leq \bar{\theta}_{ab} & \forall\, (a,b) \in N \\
\mu_p \geq 0 & \forall\, p \in P \\
\lambda_a \geq 0 & \forall\, a \in A \\
\bar{\lambda}_{ab} \geq 0, \bar{\lambda}_{ba} \geq 0, \lambda_{ab} \geq 0 & \forall\, (a,b) \in N
\end{cases}
\end{aligned}
$$

The dual to this linear program is given by

$$
\begin{aligned}
&\min \quad \sum_{a \in A} \bar{\theta}_a x_a + \sum_{(a,b) \in N} \bar{\theta}_{ab} x_{ab} \\
&\qquad \text{subject to} \begin{cases}
\sum\limits_{a \in A(p)} -x_p \geq -1 & \forall\, p \in P \\
-x_a \geq -1 & \forall\, a \in A \\
-x_a - x_b + x_{ab} \geq -1 & \forall\, (a,b) \in N \\
x_a - x_{ab} \geq 0, x_b - x_{ab} \geq 0 & \forall\, (a,b) \in N \\
x_a \geq 0 & \forall\, a \in A \\
x_{ab} \geq 0 & \forall\, (a,b) \in N
\end{cases}
\end{aligned}
$$

Thus, we finally obtain that the optimal value of the lower bound equals

$$\min \quad \sum_{a \in A} \bar{\theta}_a x_a + \sum_{(a,b) \in N} \bar{\theta}_{ab} x_{ab} \tag{21}$$

$$\text{subject to} \begin{cases} \sum_{a \in A(p)} x_p \leq 1 & \forall\, p \in P \\ 0 \leq x_a \leq 1 & \forall\, a \in A \\ x_{ab} \leq x_a, x_{ab} \leq x_b & \forall\, (a,b) \in N \\ x_{ab} \geq x_a + x_b - 1 & \forall\, (a,b) \in N \\ x_{ab} \geq 0 & \forall\, (a,b) \in N \end{cases}$$

It is easy to see that the optimal value of (21) is the same or larger than the optimal value of (18). Indeed, the only difference between (18) and (21) is that the first constraint in (21) is tighter than the corresponding constraint in (18): $\sum_{a \in A(p)} x_a \leq 1$ implies $x_a + x_b \leq 1$ for $a, b \in A(p)$, $a \neq b$, but not the other way around. (Note that the labeling $x_a = 0.5$ for $a \in A(p)$ satisfies the latter constraint but not the former, if $|A(p)| > 2$.)

Appendix C: proof of lemma 2

Consider a local subproblem $\sigma \in I$. Let σ' be a subproblem of σ, i.e. the feasibility set of σ' is contained in the feasibility set of σ: $\Omega_{\sigma'} \subseteq \Omega_\sigma$. It can be seen that adding σ' to I as another local subproblem does not affect the optimal lower bound. Indeed, it is clear that adding σ' cannot decrease the optimal bound. The optimal bound also cannot increase since for any vector $\boldsymbol{\theta'} = (\ldots, \theta^\sigma, \theta^{\sigma'}, \ldots) \in \Omega'$, where Ω' is the constraint set for the new problem, there exists vector $\boldsymbol{\theta} = (\ldots, \theta^\sigma + \theta^{\sigma'}, \ldots) \in \Omega$ whose bound is not worse since

$$\Phi_\sigma(\theta^\sigma + \theta^{\sigma'}) = 2\Phi_\sigma\left(\frac{\theta^\sigma + \theta^{\sigma'}}{2}\right) \geq \Phi_\sigma(\theta^\sigma) + \Phi_{\sigma'}(\theta^{\sigma'}).$$

(The inequality holds since $\Phi_{\sigma'}$ is the same function as Φ_σ, and it is concave.)

Let us prove part (a). Let I be a set of subproblem indexes which does not include the linear problem L. We need to show that adding L to I cannot increase the optimal lower bound. Instead of L, let us add a new subproblem p to I for each point $p \in P$ which includes only assignments in $A(p)$ (and does not include any edges), i.e. the feasibility set Ω_p for this subproblem is defined by $\theta^p_a = 0$ for all assignments $a \in A - A(p)$ and $\theta^p_{ab} = 0$ for all edges $(a,b) \in N$. As follows from the argument above and conditions of part (a), this operation cannot improve the best lower bound. Thus, it suffices to prove that replacing the new set of subproblems with L would not improve optimal bound. In other words, we need to show that for any θ^L

$$\begin{aligned} \Phi_L(\theta^L) \;\; &\leq \max \quad \sum_{p\in P} \Phi_p(\theta^p) \\ &\text{subject to} \sum_{p\in P} \theta^p = \theta^L \end{aligned} \tag{22}$$

Using LP duality, it is easy to show that in fact an equality holds in (22). Indeed, the optimal solution for vector θ^p can be obtained as follows: $\Phi_p(\theta_p) = \min\{0, \min_{a\in A(p)} \theta^p_a\}$. Thus, the maximization problem in (22) can be written as

$$\begin{aligned} &\max \quad \sum_{p\in P} -\mu_p \\ &\text{subject to} \begin{cases} \theta^p_a + \theta^q_a = \theta^L_a & \forall\, a = (p,q) \in A \\ -\mu^p \leq \theta^p_a & \forall\, p \in P, a \in A(p) \\ -\mu_p \leq 0 & \forall\, p \in P \end{cases} \end{aligned}$$

Constraints

$$\begin{aligned} \theta^p_a + \theta^q_a &= \theta^L_a \\ -\mu^p &\leq \theta^p_a \\ -\mu^q &\leq \theta^q_a \end{aligned}$$

for $a = (p,q) \in A$ can be replaced with a single constraint $-\mu^p - \mu^q \leq \theta^L_a$ since variables θ^p_a and θ^q_a are not involved in any other constraints. Then we arrive at the linear program (19) which equals $\Phi_L(\theta^L)$.

Let us now prove part (b). Using a similar argumentation, we conclude that it suffices to prove that

$$\begin{aligned} \Phi_M(\theta^M) \leq \max \quad &\sum_{a\in A} \Phi_a(\theta^a) + \sum_{(a,b)\in N} \Phi_{ab}(\theta^{ab}) \\ &\text{subject to} \sum_{a\in A} \theta^a + \sum_{(a,b)\in N} \theta^{ab} = \theta^M \end{aligned} \tag{23}$$

where $\sigma = a$ is a local subproblem in which only the element θ^a_a is allowed to be non-zero and $\sigma = (a,b)$ is a local subproblem in which only the elements θ^{ab}_a, θ^{ab}_b, θ^{ab}_{ab} are allowed to be non-zero.

It can be shown that if we take $\Phi_{ab}(\theta^{ab})$ to be a lower bound $\min_{\boldsymbol{x}\in\{0,1\}^A} E(\boldsymbol{x}|\theta^{ab})$ rather than the global minimum $\min_{\boldsymbol{x}\in M} E(\boldsymbol{x}\,|\,\theta^{ab})$ then we get an equality in 23. (An equivalent fact was proved in [37].) This implies (23) since using the global minimum instead of a lower bound can only increase the RHS.

For completeness, let us prove this equality. We have

$$\begin{aligned} \Phi_a(\theta^a) &= \min\{0, \theta^a_a\} \\ \Phi_{ab}(\theta^{ab}) &= \min\{0, \theta^{ab}_a, \theta^{ab}_b, \theta^{ab}_a + \theta^{ab}_b + \theta^{ab}_{ab}\} \end{aligned}$$

Thus, the maximization problem in (23) can be written as

$$\max \quad \sum_{a\in A} -\lambda_a + \sum_{(a,b)\in N} -\lambda_{ab}$$

$$\text{subject to} \begin{cases} \theta_a^a + \sum_{(a,b)\in N} \theta_a^{ab} = \theta_a^M & \forall\, a \in A \\ -\lambda_a \le \theta_a^a & \forall\, a \in A \\ -\lambda_a \le 0 & \forall\, a \in A \\ -\lambda_{ab} \le \theta_a^{ab}, -\lambda_{ab} \le \theta_b^{ab} & \forall\, (a,b) \in N \\ -\lambda_{ab} \le \theta_a^{ab} + \theta_b^{ab} + \theta_{ab}^M & \forall\, (a,b) \in N \\ -\lambda_{ab} \le 0 & \forall\, (a,b) \in N \end{cases}$$

We can eliminate θ_a^a from the first and the second constraint and combine them into one constraint, then we obtain

$$\max \quad \sum_{a\in A} -\lambda_a + \sum_{(a,b)\in N} -\lambda_{ab}$$

$$\text{subject to} \begin{cases} -\lambda_a + \sum_{(a,b)\in N} \theta_a^{ab} \le \theta_a^M & \forall\, a \in A \\ \lambda_a \ge 0 & \forall\, a \in A \\ \theta_a^{ab} + \lambda_{ab} \ge 0, \theta_b^{ab} + \lambda_{ab} \ge 0 & \forall\, (a,b) \in N \\ -\lambda_{ab} - \theta_a^{ab} - \theta_b^{ab} \le \theta_{ab}^M & \forall\, (a,b) \in N \\ \lambda_{ab} \ge 0 & \forall\, (a,b) \in N \end{cases}$$

Let us use variables $\bar{\lambda}_{ab}$ instead of θ_a^{ab} such that $\theta_a^{ab} = \bar{\lambda}_{ab} - \lambda_{ab}$, or $\bar{\lambda}_{ab} = \theta_a^{ab} + \lambda_{ab}$. It is straightforward to see that then we arrive at the linear program (20) which equals $\Phi_M(\theta^M)$.

Acknowledgements. We are grateful to Timothee Cour, Praveen Srinivasan, and Jianbo Shi for providing the software of their SMAC algorithm. We thank John Duchi, Gal Elidan and Danny Tarlow for sharing code and answering questions about the COMPOSE method. Thanks to Tiberio Caetano for providing the Hotel feature data.

References

1. Belhumeur, P.N.: A binocular stereo algorithm for reconstructing sloping, creased, and broken surfaces in the presence of half-occlusion. In: ICCV. (May 1993)
2. Kolmogorov, V., Zabih, R.: Computing visual correspondence with occlusions using graph cuts. In: ICCV. (2001)
3. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV. (2004)
4. Dorko, G., Schmid, C.: Selection of scale-invariant parts for object class recognition. In: ICCV. (2003)

5. Grauman, K., Darrell, T.: The pyramid match kernel: discriminative classification with sets of image features. In: ICCV. (2005)
6. Sivic, J., Russell, B., Efros, A., Zisserman, A.: Discovering object categories in image collections. In: ICCV. (2005)
7. Wallraven, C., Caputo, B., Graf, A.: Recognition with local features: the kernel recipe. In: ICCV. (2003)
8. Willamowski, J., Arregui, D., Csurka, G., Dance, C., Fan, L.: Categorizing nine visual classes using local appearance descriptors. In: Workshop on Learning for Adaptable Visual Systems, ICPR. (2004)
9. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classifcation of texture and object categories: An in-depth study. Technical Report RR-5737, INRIA Rhone-Alpes (November 2005)
10. Caetano, T.S., Cheng, L., Le, Q.V., Smola, A.J.: Learning graph matching. In: ICCV. (2007)
11. Bertsekas, D.: Nonlinear Programming. Athena Scientific (1999)
12. Chardaire, P., Sutter, A.: A decomposition method for quadratic zero-one programming. Management Science **41**(4) (1995) 704–712
13. Wainwright, M., Jaakkola, T., Willsky, A.: MAP estimation via agreement on trees: Message-passing and linear-programming approaches. IEEE Trans. Information Theory **51(11)** (2005)
14. Schlesinger, M.I., Giginyak, V.V.: Solution to structural recognition (MAX,+)-problems by their equivalent transformations. Part 1. Control Systems and Computers (1) (2007) 3–15
15. Schlesinger, M.I., Giginyak, V.V.: Solution to structural recognition (MAX,+)-problems by their equivalent transformations. Part 2. Control Systems and Computers (2) (2007) 3–18
16. Komodakis, N., Paragios, N., Tziritas, G.: MRF optimization via dual decomposition: Message-passing revisited. In: ICCV. (2007)
17. Cour, T., Srinivasan, P., Shi, J.: Balanced graph matching. In: NIPS. (2007)
18. Duchi, J., Tarlow, D., Elidan, G., Koller, D.: Using combinatorial optimization within max-product belief propagation. In: NIPS. (2007)
19. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. PAMI **24**(4) (2002) 509–522
20. Torr, P.H.S.: Geometric motion segmentation and model selection. Philosophical Transactions of the Royal Society (1998) 1321–1340
21. Metaxas, D., Koh, E., Badler, N.: Multi-level shape representation using global deformations and locally adaptive finite elements. IJCV **25**(1) (1997) 49–61
22. Sclaroff, S., Pentland, A.: Modal matching for correspondence and recognition. PAMI **17**(6) (1997) 545–561
23. Berg, A., Berg, T., Malik, J.: Shape matching and object recognition using low distortion correspondence. In: CVPR. (2005)
24. Leordeanu, M., Hebert, M.: A spectral technique for correspondence problems using pairwise constraints. In: ICCV. (2005)
25. Torr, P.H.S.: Solving Markov random fields using semi definite programming. In: AISTATS. (2003)
26. Liu, C.K., Hertzmann, A., Popović, Z.: Learning physics-based motion style with nonlinear inverse optimization. ACM Trans. on Gr. **24**(3) (2005) 1071–1081
27. Conte, D., Foggia, P., Sansone, C., Vento, M.: Thirty years of graph matching in pattern recognition. Int. J. of Pattern Recognition and Artificial Intelligence **18**(3) 265–298
28. Gold, S., Rangarajan, A.: A graduated assignment algorithm for graph matching.PAMI**18**(4) (1996) 377–388
29. Maciel, J., Costeira, J.: A global solution to sparse correspondence problems. PAMI **25**(2) (2002) 187–199
30. Cabot, A., Francis, R.: Solving certain nonconvex quadratic minimization problems by ranking the extreme points. Operations Research **18**(1) (1970) 82–86

31. Schellewald, C., Schnörr, C.: Probabilistic subgraph matching based on convex relaxation. In: EMMCVPR. (2005)
32. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. PAMI **28(10)** (October 2006) 1568–1583
33. Werner, T.: A linear programming approach to max-sum problem: A review. PAMI **29(7)** (2007)
34. Shor, N.Z.: Minimization methods for nondifferentiable functions. Springer-Verlag (1985)
35. Ahuja, R., Magnanti, T., Orlin, J.: Network Flows: Theory, Algorithms, and Applications. Prentice Hall (1993)
36. Boros, E., Hammer, P.: Pseudo-boolean optimization. Discr. Appl. Math. **123(1-3)** (2002)
37. Hammer, P.L., Hansen, P., Simeone, B.: Roof duality, complementation and persistency in quadratic 0-1 optimization. Mathematicl Programming **28** (1984) 121–155
38. Boros, E., Hammer, P.L., Sun, X.: Network flows and minimization of quadratic pseudo-Boolean functions. Technical Report RRR 17-1991, RUTCOR (May 1991)
39. Nedic, A., Bertsekas, D.: Incremental subgradient methods for nondifferentiable optimization. SIAM J. Optimization **12(1)** (2001) 109–138
40. Lim, C., Sherali, H.D.: Convergence and computational analyses for some variable target value and subgradient deflection methods. Computational Optimization and Applications **34**(3) (2006) 409–428
41. Lempitsky, V., Rother, C., Blake, A.: LogCut - efficient graph cut optimization for Markov random fields. In: ICCV. (2007)
42. Rother, C., Kolmogorov, V., Lempitsky, V., Szummer, M.: Optimizing binary MRFs via extended roof duality. In: CVPR. (2007)
43. Elidan, G., McGraw, I., Koller, D.: Residual belief propagation: Informed scheduling for asynchronous message passing. In: UAI. (2006)
44. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11) (1998) 2278–2324
45. Boros, E., Hammer, P.L., Tavares, G.: Preprocessing of unconstrained quadratic binary optimization. Technical Report RRR 10-2006, RUTCOR (April 2006)