Identity inference: generalizing person re-identification scenarios

Svebor Karaman and Andrew D. Bagdanov

Media Integration and Communication Center
University of Florence, Viale Morgagni 65, Florence, Italy
svebor.karaman@unifi.it,bagdanov@dsi.unifi.it

Abstract. In this article we introduce the problem of identity inference as a generalization of the re-identification problem. Identity inference is applicable in situations where a large number of unknown persons must be identified without knowing *a priori* that groups of test images represent the same individual. Standard single- and multi-shot person re-identification are special cases of our formulation. We present an approach to solving identity inference problems using a Conditional Random Field (CRF) to model identity inference as a labeling problem in the CRF. The CRF model ensures that the final labeling gives similar labels to detections that are similar in feature space, and is flexible enough to incorporate constraints in the temporal and spatial domains. Experimental results are given on the ETHZ dataset. Our approach yields state-of-the-art performance for the multi-shot re-identification task and promising results for more general identity inference problems.

1 Introduction

Person re-identification is traditionally defined as the recognition of an individual at different times, over different camera views and/or locations, and considering a large number of candidate individuals. It is a standard component of multi-camera surveillance systems as it is a way to associate multiple observations of the same individual over time. Particularly in scenarios in which the long-term behavior of persons must be characterized, accurate re-identification is essential. In realistic, wide-area surveillance scenarios such as airports, metro and train stations, re-identification systems should be capable of robustly associating a unique identity with hundreds, if not thousands, of individual observations.

Re-identification performance is usually evaluated as a retrieval problem. Given a gallery consisting of a number of known individuals and images of each, for each test image or group of test images of an unknown person the goal of re-identification is to return a ranked list of individuals from the gallery. Configurations of the re-identification problem are generally classified according to how much group structure is available in the gallery and test image sets. In a single-shot image set there is no grouping information availabled. Though there might be multiple images of an individual, there is no knowledge of which images

correspond to that person. In a multi-shot image set, on the other hand, there is explicit grouping information available. That is, it is known which images correspond to the same individual.

The classification of re-identification scenarios into multi- and single-shot configurations is useful for establishing benchmarks and standardized datasets for experimentation on person re-identification. However, these scenarios are not particularly realistic with respect to some real-world application scenarios. In video surveillance scenarios, for example, it is more common to have a few individuals of interest and to desire that all occurrences of those individuals be labeled. In this case the number of unlabeled test images to re-identify is typically much larger than the number of gallery images available.

In this article we propose a generalization of person re-identification which we call *identity inference*. The identity inference formulation is expressive enough to represent existing single- and multi-shot scenarios, while at the same time also modeling a larger class of problems not discussed in the literature. We also propose a CRF-based approach to solving identity inference problems. Our model is able to efficiently and accurately solve a broad range of identity inference problems, including existing person re-identification scenarios as well as more difficult tasks involving very many unlabeled test images.

2 Related work

The majority of existing research on the person re-identification problem has concentrated on the development of sophisticated features for describing the visual appearance of targets. In [1] were introduced discriminative appearance-based models using Partial Least Squares (PLS) over texture, gradients and color features. The authors of [2] use an ensemble of local features learned using a boosting procedure, while in [3] the authors use a covariance matrix of features computed in a grid of overlapping cells. The SDALF descriptor introduced in [4] exploits axis symmetry and asymmetry and represents each part of a person by a weighted color histogram, maximally stable color regions and texture information from recurrent high-structured patches. In [5] the authors fit a Custom Pictorial Structure (CPS) model consisting of head, chest, thighs and legs part descriptors using color histograms and Maximally Stable Color Region (MSCR). The Global Color Context (GCC) of [6] uses a quantization of color into color words and then builds a color context modeling the self-similarity for each word using a polar grid. The Histogram Plus Epitome (HPE) approach in [7] represents a person by a global mean color histogram and recurrent local patterns through epitomic analysis.

The approaches mentioned above concentrate on feature representation and not specifically on the classification or ranking technique. And approach which does concentrate specifically on the ranking approach is the Ensemble RankSVM technique of [8], which learns a ranking SVM model to solve the single-shot re-identification problem.

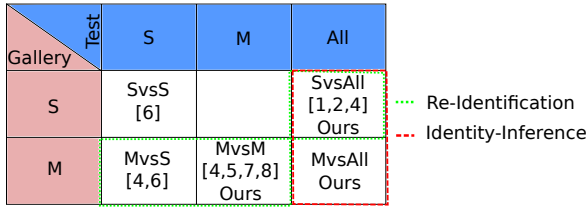


Fig. 1: Re-identification and identity-inference protocols. Though the authors of [6] use "single" to describe their test sets, they only use one image per person.

We believe that in realistic scenarios many unlabeled images will be available while only few detections with known identities will be given, which is a scenario not covered by the standard classification of single- and multi-shot cases. We propose a CRF model that is able to encode a "soft grouping" property of unlabeled images. Our application of CRFs to identity inference is similar in spirit to recent work using CRFs for multi-target tracking [9]. However, to the best of our knowledge CRFs have not been directly applied to the re-identification problem and in this article we explore their use on identity inference as formulated in next section.

3 Identity inference as generalization of re-identification

In this section we give a formal definition of the re-identification and identity inference problems. The literature on person re-identification covers many different configurations of gallery and test images. We consider each in turn and show how each can be represented as an instance of our definition of re-identification. A summary of the different protocols with corresponding works is given in figure 1.

Let $\mathcal{L} = \{1, \dots N\}$ be a label set for a re-identification scenario, where each element represents a unique individual appearing in a video sequence or collection of sequences. We assume that there are a number of instances (images) of individuals from $\mathcal{L}$ detected in a video collection:

$$\mathcal{I} = \{x_i \mid i = 1 \dots D\}\,.$$

We assume that each image x_i of an individual is represented by a feature vector $\mathbf{x}_i \equiv \mathbf{x}(x_i)$ and that the label corresponding to instance x_i is given by $y_i \equiv y(x_i)$.

An instance of a re-identification problem, represented as a tuple $\mathcal{R} = (\mathcal{X}, \mathcal{Z})$, is completely characterized by its gallery and test image sets ($\mathcal{X}$ and $\mathcal{Z}$, respectively). Formally, the gallery images are defined as:

$$\mathcal{X} = \{\mathcal{X}_j \mid j = 1 \dots N\}\,, \text{where } \mathcal{X}_j \subset \{x \mid y(x) = j\}\,.$$

That is, for each individual i, a subset of all available images is chosen to form his gallery: $\mathcal{X}_i$. The set of test images is defined as:

$$\mathcal{Z} = \{\mathcal{Z}_j \mid j = 1 \dots M\} \subset \mathcal{P}(\mathcal{I}),$$

where $\mathcal{P}$ is the powerset operator (i.e. $\mathcal{P}(I)$ is the set of all subsets of $\mathcal{I}$). We further require for all $\mathcal{Z}_j \in \mathcal{Z}$ that $x, x' \in \mathcal{Z}_j \Rightarrow y(x) = y(x')$ (sets in $\mathcal{Z}$ have homogeneous labels), and $\mathcal{Z}_j \in \mathcal{Z} \Rightarrow \mathcal{Z}_j \cap \mathcal{X}_i = \emptyset, \forall i \in \{1 \dots N\}$ (the test and gallery sets are disjoint). A solution to an instance of a re-identification problem is a mapping from the test images $\mathcal{Z}$ to the set of all permutations of $\mathcal{L}$.

3.1 Re-identification scenarios

Single-versus-all re-identification (SvsAll) is often referred to as simply *single-shot re-identification* or *single-versus-single* (SvsS) but could better be described as *single-versus-all* (SvsAll) [1] re-identification, see figure 1. In the SvsAll re-identification scenario a single gallery images is given for each individual, and *all remaining instances* of each individual are used for testing: $M = D - N$. Formally, a single-versus-all re-identification problem is a tuple $\mathcal{R}_{\text{SvsAll}} = (\mathcal{X}, \mathcal{Z})$, where:

$$\begin{aligned} \mathcal{X}_j &= \{x\} \text{ for some } x \in \{x \mid y(x) = j\}, \text{ and} \\ \mathcal{Z}_j &= \{\{x\} \mid x \in \mathcal{I} \setminus \mathcal{X}_j \text{ and } y(x) = j\} \end{aligned}$$

Multi-versus-single shot re-identification (MvsS) is defined using G gallery images of each person, while each of the test sets $\mathcal{Z}_j$ contains only a single image. In this case $M = N$, as there are exactly as many partial test sets $\mathcal{Z}_j$ as persons depicted in the gallery. Formally, a MvsS re-identification problem is a tuple $\mathcal{R}_{\text{MvsS}} = (\mathcal{X}, \mathcal{Z})$, where:

$$\begin{aligned} \mathcal{X}_j &\subset \{x \mid y(x) = j\} \text{ and } |\mathcal{X}_j| = G \ \forall j \text{ and} \\ \mathcal{Z}_j &= \{x\} \text{ for some } x \notin \mathcal{X}_j \text{ s.t. } y(x) = j. \end{aligned}$$

The MvsS configuration is not precisely a generalization of the SvsAll person re-identification problem in that, after selecting G gallery images for each individual, only a *single* test image is selected to form the test sets $\mathcal{Z}_j$.

Multi-versus-multi shot re-identification (MvsM) is the case in which the gallery and test sets of each person both have G images. Formally, a MvsM re-identification problem is a tuple $\mathcal{R}_{\text{MvsM}} = (\mathcal{X}, \mathcal{Z})$, where:

$$\begin{aligned} \mathcal{X}_j &\subset \{x \mid y(x) = j\} \text{ and } |\mathcal{X}_j| = G \ \forall j \text{ and} \\ \mathcal{Z}_j &\subset \{x \mid y(x) = j \text{ and } x \notin \mathcal{X}_j\} \text{ and } |\mathcal{Z}_j| = G \ \forall j. \end{aligned}$$

The goal in MvsM re-identification is to re-identify each *group* of test images, leveraging the knowledge that images in each group are all of the same individual.

[1] We use the SvsAll terminology as the SvsS terminology could be (mis-)interpreted as in [6] in which only a *single* image was selected for each individual in the test set.

3.2 Identity inference

Identity inference addresses the problem of having *few* labeled images while desiring to label *many* unknown images without explicit knowledge that groups of images represent the same individual. The formulation of the *single-versus-all* re-identification falls within the scope of identity inference, but neither the multi-versus-single nor the multi-versus-multi formulations are a generalization of this case to multiple gallery images. In the MvsS and MvsM cases the test set is either a singleton for each person (MvsS) or a group of images (MvsM) of the same size as the gallery image set for each person. Identity inference could be described as a *multi-versus-all* configuration. Formally, it is a tuple $\mathcal{R}_{\text{MvsAll}} = (\mathcal{X}, \mathcal{Z})$, where:

$$\begin{aligned} &\mathcal{X}_j \subset \{x \mid y(x) = j\} \text{ and } |\mathcal{X}_j| = G \text{ and} \\ &\mathcal{Z}_j = \{\{x\} \mid x \in \mathcal{I} \setminus \mathcal{X}_j \text{ and } y(x) = j\} \end{aligned}$$

In instances of identity inference a set of G gallery images are chosen for each individual. Each remaining images of each individual is then used as an element of the test set without any identity grouping information. As in the SvsAll case, the test images sets are all singletons.

4 A CRF model for identity inference

Conditional Random Fields (CRFs) have been used to model the statistical structure of problems such as semantic image segmentation [10], and stereo matching [11]. In this section we show how we model the identity inference problem as a minimum energy labeling problem in a CRF.

A CRF is defined by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a set of random variables $\mathcal{Y} = \{Y_j \mid j = 1 \dots |V|\}$ which represent the statistical structure of the problem being modelled, and a set of possible labels $\mathcal{L}$. The vertices $\mathcal{V}$ index the random variables in $\mathcal{Y}$ and the edges $\mathcal{E}$ encode the statistical dependence relations between the random variables. The labeling problem is then to find an assignment of labels to nodes that minimizes an energy function E over possible labelings $\mathbf{y}^* = (y_i^*)_{i=1}^{|V|}$: $\tilde{\mathbf{y}} = \arg\min_{\mathbf{y}^*} E(\mathbf{y}^*)$. The energy function $E(\mathbf{y}^*)$ is defined as:

$$E(\mathbf{y}^*) = \sum_{i \in \mathcal{V}} \phi_i(y_i^*) + \lambda \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(y_i^*, y_j^*), \tag{1}$$

where $\phi_i(y_i^*)$ is a unary data potential encoding the cost of assigning label y_i^* to vertex i and $\psi_{ij}(y_i^*, y_j^*)$ is a binary smoothness potential representing the conditional cost of assigning labels y_i^* and y_j^* respectively to vertices i and j. The parameter λ in equation (1) controls the tradeoff between data and smoothness costs. Given an instance of a CRF, there exist very efficient algorithms for finding the optimal labeling $\tilde{\mathbf{y}}$ using, for example, graph cuts [12, 13].

We can map an identity inference problem $\mathcal{R} = (\mathcal{X}, \mathcal{Z})$ onto a CRF by defining the vertex and edge sets $\mathcal{V}$ and $\mathcal{E}$ in terms of the gallery and test image

sets defined by $\mathcal{X}$ and $\mathcal{Z}$. We have found two configurations of vertices and edges to be useful for solving identity inference problems. The first uses vertices to represent groups of images in the test set $\mathcal{Z}$ and is particularly useful for modeling MvsM re-identification problems:

$$\mathcal{V} = \bigcup_{i=1}^{N} \mathcal{Z}_i \text{ and } \mathcal{E} = \{(x_i, x_j) \mid x_i, x_j \in \mathcal{Z}_l \text{ for some } l\}\,.$$

The edge topology in this CRF is completely determined by the group structure as expressed by the $\mathcal{Z}_j$.

When no identity grouping information is available for the test set, as in the general identity inference case as well as in SvsAll re-identification, we instead use the following formulation of the CRF:

$$\mathcal{V} = I \text{ and } \mathcal{E} = \bigcup_{x_i \in \mathcal{V}} \{(x_i, x_j) \mid x_j \in \text{kNN}(x_i)\}\,,$$

where the $\text{kNN}(x_i)$ maps an image to its k most similar images in feature space. The topology of this CRF formulation, in the absence of explicit group information, uses feature similarity to form connections between nodes.

The unary data potential determines the cost of assigning label y_i^* to vertex i given $\mathbf{x}(x_i)$, the observed feature representation of image x_i. It is proportional to the minimum L1-distance between the feature representation of image x_i and any gallery image of individual y_i^*. We define it as:

$$\phi_i(y_i^*) = \begin{cases} 1 & \text{if } x_i \in \mathcal{X} \text{ and } y_i^* \neq y(x_i) \\ \min_{x \in \mathcal{X}_{y_i^*}} ||\mathbf{x}(x) - \mathbf{x}(x_i)|| & \text{otherwise.} \end{cases}$$

If a vertex corresponds to a gallery image, its data potential is 1 for every incorrect label and zero for the correct one.

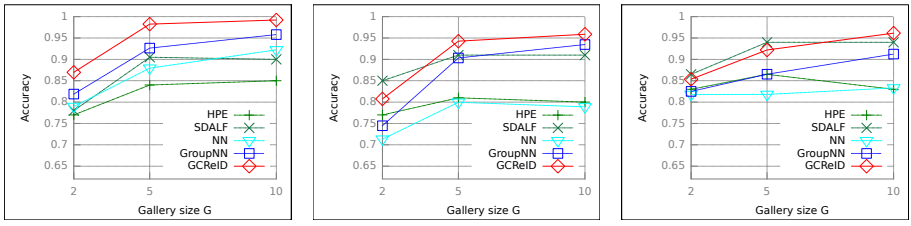
Without explicit neighborhood topology given by identity groups, we use the smoothness potential to encourage similar detections to share the same labels:

$$\psi_{ij}(y_i^*, y_j^*) = w_{ij} \min_{\substack{x \in \mathcal{X}_{y_i^*} \\ x' \in \mathcal{X}_{y_j^*}}} ||\mathbf{x}(x) - \mathbf{x}(x')||. \tag{2}$$

This smoothness potential ensures local consistency between labels in neighboring nodes: the more similar two labels are in terms of the available gallery images for them, the lower the cost for them to be labeled the same in the CRF. The weighting factors w_{ij} allow the smoothness potential between nodes i and j to be flexibly controlled according to the problem at hand.

5 Experimental results

For evaluating identity inference we are especially interested in test scenarios where there are many unlabeled images of each test subject. For this reason, we



(a) ETHZ1 (b) ETHZ2 (c) ETHZ3

Fig. 2: MvsM re-identification accuracy. Note that these are *not* CMC curves, but are Rank-1 *classification* accuracies over varying gallery and test set sizes.

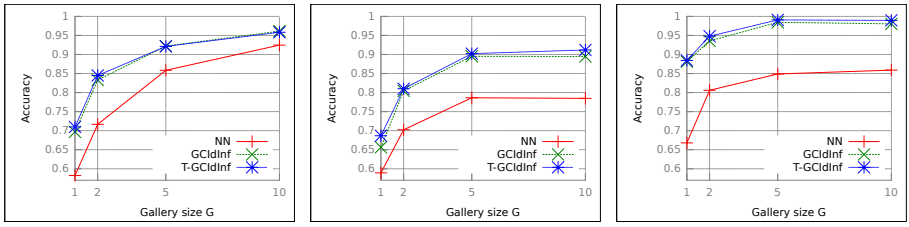
selected the publicly available ETHZ [1] dataset which consists of three video sequences, because on average each person appears in more than 50 images. This dataset is also interesting because the frame number of each detection is available and can therefore be incorporated into our CRF model as a temporal constraint.

As feature representation we compute Hue-SIFT [14] features on a dense grid, quantize them to a vocabulary of 512 visual words and group them into cells of a 1×6 cell spatial pyramid [15]. In all experiments on a specific ETHZ sequence, the codebook for the visual vocabulary was learned through k-means clustering on features from the remaining two ETHZ sequences. The choice of six horizontal stripes for the spatial pyramid representation is similar to the choice made in [8]. We should note that the CRF model we propose is orthogonal to the choice of feature descriptor and most descriptors discussed in the literature could be used in our framework.

5.1 Re-identification

Here we apply our CRF framework to solve MvsM re-identification problems. In these experiments we fix $\lambda = 1$ in the energy function of equation (1) and evaluate performance for galleries varying in size over 2, 5 and 10 images per person. For each configuration, we randomly select the gallery and test images and average performance over ten trials. Note that grouping information in the test set is explicitly encoded in the CRF: edges only link test images that correspond to the same individual. Results on MvsM person re-identification are presented in figure 2. We compare our results, which we refer to as GCReID for "Graph Cut Re-Identification", with the published results of SDALF [4] and HPE [7]. The NN curve in figure 2 corresponds to labeling each test image with the nearest gallery image label without exploiting group knowledge, while the GroupNN approach use this knowledge by setting the label of the group of test images as the label of the model which distance to any of these test images is minimal.

The results in figure 2 show that, by using the CRF we can ensure a more consistent labeling especially when having a higher number of test images, thus outperforming state-of-the-art methods even though our descriptor is less sophis-



(a) ETHZ1 (b) ETHZ2 (c) ETHZ3

Fig. 3: Identity inference accuracy on ETHZ datasets.

ticated. The use of the CRF to enforce labeling consistency allows our approach to outperform simpler, ad hoc reasoning about group similarity (see the results of the GroupNN method). We also note that, while the approach of SDALF [4] computes an accumulated single descriptor from multiple gallery images, we keep the multiple appearances in our model. This, in combination with the inference in our CRF framework, enables us to obtain extremely good results on the ETHZ1 dataset (figure 2a). While other approaches in the literature tend to have lower results with a growing number of persons our approach seems to be more robust in these situations.

5.2 Identity inference

In the CRF model proposed in section 4 for identity inference in which no identity grouping information is available for the test image sets, the local neighborhood structure is determined by the K nearest neighbors to each image in feature space. For all experiments we set $K = 4$. For the general identity inference case, unlike MvsM person re-identification, we have no information about relationships between test images. We define the weights w_{ij} from equation (2) between vertices i and j in the CRF in terms of feature similarity and a temporal constraint:

$$w_{ij} = (1 - \alpha)(1 - ||\mathbf{x}(x_i) - \mathbf{x}(x_j)||) + \alpha\tau_{ij}, \tag{3}$$

where $\alpha \in [0, 1]$ is a weighting factor controlling the tradeoff between temporal and feature similarities, τ_{ij} is a temporal weighting factor defined as:

$$\tau_{ij} = \begin{cases} 1 - \frac{|f_i - f_j|}{\tau} & \text{if } |f_i - f_j| \leq \tau \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where f_i and f_j are the frame numbers in which detections i and j occurred, respectively, and τ is a threshold limiting the temporal influence to a finite number of frames. In preliminary experiments we found $\tau = 25$ to work well for a variety of configurations and we use this temporal window in all of our reported results. Similarly, we use a value of $\alpha = 0.3$ which places more attention on similarity in feature space. The results in figure 3 are given for λ set to 5.

Fig. 4: Identity inference results (SvsAll). First row: test image, second row: incorrect NN result, third row: correct result given by GCIdInf.

In identity inference, gallery images are randomly selected and *all remaining images* define the test set. Results are averaged over 10 trials as before. Using our CRF framework clearly improves accuracy over the simple NN model. The best configuration is T-GCIDInf which uses feature similarity-weighted edges with temporal constraints, yielding an average improvement of 15% over the three datasets with respect to nearest-neighbor labeling. Our approach permits us to label a large number of unknown images using only few gallery images for each person. For example, on the ETHZ3 dataset we are able to correctly label 1596 out of 1706 test images using only 2 model images per person. The robustness of our method with respect to occlusions and illumination changes is shown in the qualitative results shown in figure 4. The CRF approach proposed yields correct labels even in strongly occluded cases thanks to the neighborhood edges connecting it to less occluded, yet similar, images.

6 Discussion

In this paper we have introduced the identity inference problem which we propose as a generalization of the standard person re-identification problem described in the literature. Identity inference can be thought of as a generalization of the single-versus-all shot case of person re-identification, and at the same time as a relaxation of the multi-versus-multi shot case. Instances of identity inference problems do not require hard knowledge about relationships between test images (e.g. that they correspond to the same individual). We have also proposed a CRF-based approach to solving identity inference problems. Our solution uses feature space and temporal (when available) similarity to define the neighborhood topology in the CRF. Our experimental results show that the CRF approach can efficiently solve standard re-identification tasks, achieving

classification performance comparable to state-of-the-art Rank-1 results in the literature. The CRF model can also be used to solve more general identity inference problems in which no hard grouping information and very many test images are present in the test set. Our current work concentrates on exploring more powerful descriptors and more realistic configurations for identity inference in the real world. To this end we are also working on developing a multi-camera dataset for identity inference.

References

1. Schwartz, W., Davis, L.: Learning discriminative appearance-based models using partial least squares. In: Proceedings of SIBGRAPI, IEEE (2009) 322–329
2. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proceedings of ECCV. (2008) 262–275
3. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Multiple-shot human re-identification by mean riemannian covariance grid. In: Proceedings of AVSS. (2011) 179–184
4. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. Proceedings of CVPR (2010) 2360–2367
5. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: Procedings of BMVC. (2011)
6. Cai, Y., Pietikäinen, M.: Person re-identification based on global color context. In: Proceedings of ACCV Workshops. (2011) 205–215
7. Bazzani, L., Cristani, M., Perina, A., Murino, V.: Multiple-shot person re-identification by chromatic and epitomic analyses. Pattern Recognition Letters **33** (2012) 898–903
8. Prosser, B., Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: Proceedings of BMVC. (2010)
9. Yang, B., Nevatia, R.: An online learned crf model for multi-target tracking. In: Proceedings of CVPR. (2012)
10. Boix, X., Gonfaus, J., van de Weijer, J., Bagdanov, A., Serrat, J., Gonzàlez, J.: Harmony potentials. International Journal of Computer Vision (2012) 1–20
11. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision **47** (2002) 7–42
12. Kolmogorov, V., Zabin, R.: What energy functions can be minimized via graph cuts? IEEE Transactions on Pattern Analysis and Machine Intelligence **26** (2004) 147–159
13. Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., Rother, C.: A comparative study of energy minimization methods for markov random fields with smoothness-based priors. IEEE Transactions on Pattern Analysis and Machine Intelligence **30** (2008) 1068–1080
14. van de Weijer, J., Schmid, C.: Coloring local feature extraction. In: Proceedings of ECCV. Volume Part II. (2006) 334–348
15. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proceedings of CVPR. Volume 2., Ieee (2006) 2169–2178