Orientation Invariant Features for Multiclass Object Recognition

Michael Villamizar[1], Alberto Sanfeliu[1] and Juan Andrade-Cetto[2]

[1] Institut de Robòtica i Informàtica Industrial, UPC-CSIC
Llorens Artigas 4-6, 08028 Barcelona, Spain
{mvillami,sanfeliu}@iri.upc.edu
[2] Computer Vision Center, Universitat Autònoma de Barcelona
Edifici O, Campus UAB, 08193 Bellaterra, Spain
cetto@cvc.uab.es

Abstract. We present a framework for object recognition based on simple scale and orientation invariant local features that when combined with a hierarchical multiclass boosting mechanism produce robust classifiers for a limited number of object classes in cluttered backgrounds. The system extracts the most relevant features from a set of training samples and builds a hierarchical structure of them. By focusing on those features common to all trained objects, and also searching for those features particular to a reduced number of classes, and eventually, to each object class. To allow for efficient rotation invariance, we propose the use of non-Gaussian steerable filters, together with an Orientation Integral Image for a speedy computation of local orientation.

1 Introduction

Object detection is a fundamental issue in most computer vision tasks; particularly, in applications that require object recognition. Early approaches to object recognition are based on the search for matches between user-generated geometrical object models and image features. To overcome the need of such models, appearance-based object recognition gained popularity in the past two decades using dimensionality reduction techniques such as PCAs for whole-image matching. Unfortunately, appearance based matching as such, is prone to fail in situations with modest occlusions or under varying backgrounds. Lately, a new paradigm for object recognition has appeared based on the matching of geometrical as well as appearance local features. The most popular of these, perhaps, the SIFT descriptor [1].

Financial support to M. Villamizar and A. Sanfeliu comes from the EURON Network Robot Systems Research Atelier NoE-507728, and the Spanish Ministry of Education and Science project NAVROB DPI 2004-05414. J. Andrade-Cetto is a Juan de la Cierva Postdoctoral Fellow of the Spanish Ministry of Education and Science under project TIC2003-09291, and is also funded in part by the EU PACO-PLUS project FP6-2004-IST-4-27657. The authors belong to the Artificial Vision and Intelligent Systems Group funded in part by the Catalan Research Commission.

Instead of using general saliency rules for feature selection as in the case of the SIFT descriptor, the use of boosting techniques for feature selection has proven beneficial in choosing the most discriminant geometric and appearance features from training sets. Despite their power in achieving accurate recognition from trained data, early boosting mechanisms such as [2], were tailored to single class object recognition, and are not suitable for multiclass object recognition given the large amount of features that need to be trained independently for each object class. Lately however, there have been some extensions to the general idea of classfication with boosting that allow the combined training of multiple classes [3, 4]. In the computer vision doamin, Torralba *et al.* [5] proposed an extension to one such boosting algorithm (gentleboost), with the purpose of sharing features across multiples object classes so as to reduce the total number of classifiers. They called it JointBoost, and in this approach, all object classes are trained jointly, and for each possible subset of classes ($2^n - 1$ excluding the empty set), the most useful feature is selected to distinguish that subset from the background class. The process is repeated until the overall classification error reaches a minimum, or until a limit on the number of classifiers is achieved.

The type of weak classifier features used in [5] are very simple template matching masks, that would presumibly fail if sample objects are to be found at different orientations than as trained. In this work we investigate on the use of similar multiclass feature selection, but with keen interest in fast computation of orientation invariant weak classifiers [6] for multiclass rotation invariant object recognition.

In [2], Viola introduced the integral image for very fast feature evaluation. Once computed, an integral image allows the computation of Haar-like features [7] at any location or scale in real time. Unfortunately, such system is not invariant to object rotation or occlusions. Other recognition systems that might work well in cluttered scenes are based on the computation of multi-scale local features such as the previously mentioned SIFT descriptor [1]. One key idea behind the SIFT descriptor is that it incorporates canonical orientation values for each keypoint. Thus, allowing scale and rotation invariance during recognition. Even when a large number of SIFT features can be computed in real time for one single image, their correct pairing between sample and test images is performed via nearest neighbor search and generalized Hough transform voting, followed by the solution of the affine relation between views; which might end up to be a time consuming process.

Yokono and Poggio [8, 9] settle for Harris corners at various levels of resolution as interest points, and from these, they select as object features those that are most robust to Gaussian derivative filters under rotation and scaling. As Gaussian derivatives are not rotation invariant, they use steerable filters [10] to steer all the features responses according to the local gradient orientation around the interest point. In the recognition phase, the system still requires local feature matching, and iterates over all matching pairs, in groups of 6, searching for the best matching homography, using RANSAC for outlier removal. Unfortunately, the time complexity or performance of their approach was not reported.

In [6] we realized that filter response to Haar masks can be not only be computed efficiently with an integral image scheme; but also, that such masks can be approximately rotated with some simplifications of the Gaussian steerable filter. Thus, allowing for fast computation of rotation invariant filter responses as week classifiers.

In this paper, we incorporate these two ideas, multiclass boosting, and rotation invariance, for the selection of joint and specific local features to construct a hierarchical structure that allow recognizing multiples objects independently of position, scale and orientation with a reduced set of features. In our system, keypoints are chosen as those regions in the image that have the most discriminant response under convolution with a set of wavelet basis functions at several scales and orientations. Section 2 explains how the most relevant features are selected and combined to classify multiples objects. The selection is based on JointBoost, in which a hierarchical structure is composed by sets of joint and specific classifiers. A linear combination of these weak classifiers produces a strong classifier for each object class, which is used for detection. Rotation invariance is achieved by filtering with oriented basis functions. Filter rotation is efficiently computed with the aid of a steerable filter, that is, as the linear combination of basis filters, as indicated in Section 3.

During the recognition phase, sample image regions must be rotated to a trained canonical orientation, prior to feature matching. Such orientation is dictated by the peak on a histogram of gradient orientations, depicted in Section 4. Section 5 explains our proposed Orientation Integral Image for the speed of kernel orientation computation, and Section 6 presents some experiments.

2 Feature Selection

The set of local features that best discriminates an object is obtained by convolving positive sample images with a simplified set of wavelet basis function operators [7] at different scales and orientations. These filters have spatial orientation selectivity as well as frequency selectivity, and produce features that capture the contrast between regions representing points, edges, and strips, and have high response along for example, contours. The set of operators used is shown in Figure 1. Filter response is equivalent to the difference in intensity in the original image between the dark and light regions dictated by the operator. Figure 1 d) exemplifies how an object can be represented by a small set of the most useful local features.

Convolving these operators at any desired orientation is performed by steering the filter (Section 3), and fast convolution over any region of the entire image is efficiently obtained using an integral image (Section 5).

Feature selection is performed as in JointBoost [5], choosing one at a time, from the $2n-1$ subsets of the classes $c = 1...n$ (empty set excluded), the weak classifier $h(I, s)$ that best discriminates any subset s from the background class (lowest classification error). The weak classifier is defined by the parameters

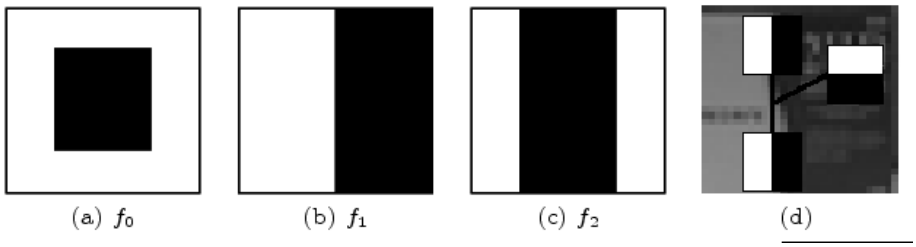


Fig. 1. Simplified wavelet basis function set. a) center-surround b) edge, and c) line; and d) object local features

filter type, size, location, orientation and threshold, taking the binary decission value

$$h(I,s) = \begin{cases} 1 & : \quad I * f > t \\ 0 & : \quad \text{otherwise} \end{cases} \tag{1}$$

where I is a training sample image of class c in the subset s, f is the filter being tested, with all its parameters, $*$ indicates the convolution operation, and t is the filter response threshold.

At each iteration during the training phase, the algorithm must find for all of the $2n-1$ subsets, the weak classifier that best discriminates that subset from the background class by minimizing the squared error over weighted samples of all classes in that subset

$$J_{wse} = \sum_{c=1}^{n} \sum_{s=1}^{m} w_i^c (z_i^c - h(I,s))^2 \tag{2}$$

where z_i^c and w_i^c are the membership label and weight of the sample i for class c respectively, and m the total number of training samples. The algorithm also updates sets of weights over the training samples. The number of sets corresponds with the number of classes to learn. Initially, all weights are set equally, but on each round, the weights of missclassified samples are increased so that the algorithm is forced to focus on such hard samples in the training set the previously chosen classifiers missed. Finally, choosing the weak classifier for the subset that had the minimum squared error J, and iteratively adding it to the Strong Classifier for every class c in s, $H(I,c)$,

$$H(I,c) := H(I,c) + h(I,s) \tag{3}$$

Scale invariance is obtained by iterating also over scaled filters within the classifier H. Scaling of the filters can be performed in constant time for a previously computed integral image.

3 Steerable Filters

In order to achieve orientation invariance, the local filters must be rotated previous to convolution. A good alternative is to compute these rotations with steerable filters [10], or with its complex version [11]. A steerable filter is a rotated filter comprised of a linear combination of a set of oriented basis filters

$$I * f(\theta) = \sum^{n} k_i(\theta) I * f(\theta_i) \ , \tag{4}$$

where $f(\theta_i)$ are the oriented basis filters, and k_i are the coefficients of the bases.

Consider for example, the Gaussian function $G(u, v) = e^{-(u^2+v^2)}$, and its first and second order derivative filters $G'_u = -2ue^{-(u^2+v^2)}$ and $G''_u = (4u^2 - 2)e^{-(u^2+v^2)}$. These filters can be re-oriented as a linear combination of filter bases. The size of the basis is one more than the derivative order.

Consequently. the first order derivative of our Gaussian function at any direction θ is

$$G'_\theta = \cos\theta G'_u + \sin\theta G'_v \ , \tag{5}$$

and, the steered 2nd order Gaussian filter can be obtained with

$$G''_\theta = \sum_{i=1}^{3} k_i(\theta) G''_{\theta_i} \tag{6}$$

with $k_i(\theta) = \frac{1}{3}(1 + 2\cos(\theta - \theta_i))$; and G''_{θ_i} precomputed second order derivative kernels at $\theta_1 = 0$, $\theta_2 = \frac{\pi}{3}$, and $\theta_3 = \frac{2\pi}{3}$. See Figure 2.

Convolving with Gaussian kernels is a time consuming process. Instead, we propose in [6] to approximate such filter response by convolving with the Haar basis with the objective of using the integral image. Thus, we approximate the oriented first derivative response with

$$I * f_1(\theta) = \cos\theta I * f_1(0) + \sin\theta I * f_1(\tfrac{\pi}{2}) \ . \tag{7}$$

and in the same sense, the filtering with our line detector at any orientation θ is obtained with

$$I * f_2(\theta) = \sum_{i=1}^{3} k_i(\theta) I * f_2(\theta_i) \ . \tag{8}$$

The similarity of the response between the Gaussian and the Haar filters allows us to use the later basis instead as weak classifiers for the detection of points, edges, and lines; just as the Gaussian filters do. The main benefit of the approach is in speed of computation. While convolution with a Gaussian kernel takes time O(n) the size of the kernel, convolution with the oriented Haar basis can be computed in constant time using an integral image representation. Figure 3 shows some results.

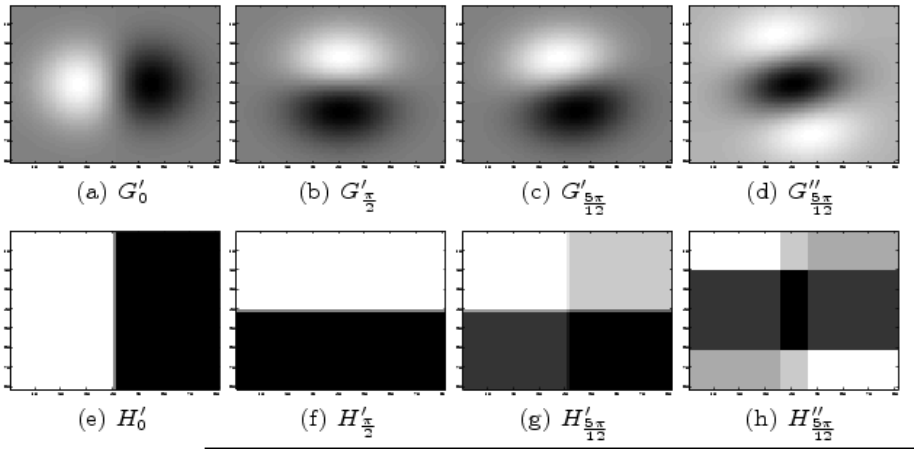


Fig. 2. First and second order steerable filters. (a-b) Gaussian basis, (c-d) Gaussian oriented filters, (e-f) Haar basis, (g-h) Haar oriented filters.

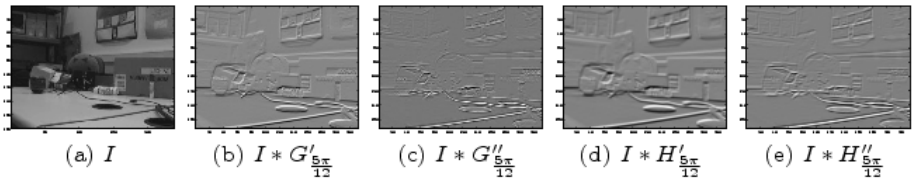


Fig. 3. Filter responses. (a) original image, (b-e) filter responses.

4 Local Orientation

Consider a training session has produced a constellation H of local features h as the one shown in Figure 4. Now, the objective is to test for multiple positions and scales in each new image, whether such constellation passes the test H or not. Instead of trying every possible orientation of our constellation, we chose to store the canonical orientation θ_0 of H from a reference training image block, and to compare it with the orientation θ of each image block being tested. The difference between the two indicates the amount we must re-orient the entire feature set before the test H is performed.

On way to compute block image orientation is with ratio of first derivative Gaussians G'_u and G'_v [9], $\tan\theta = \frac{I*G'_v}{I*G'_u}$. Another technique, more robust to partial occlusions, is to use the mode of the local gradient orientation histogram (see Figure 4 c-d), for which it is necessary to compute gradient orientations pixel by pixel, instead of a region convolution as in the previous case.

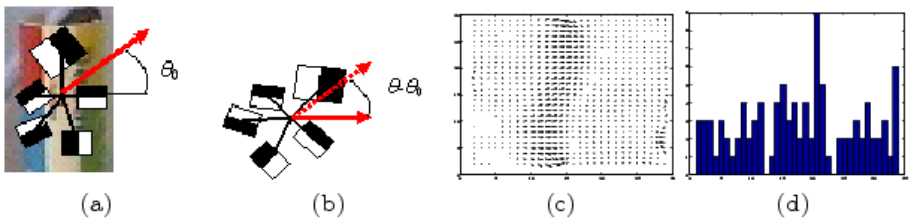


Fig. 4. Local orientation a) canonical orientation, b) rotated constellation, c) image gradients, b) gradient orientation histogram.

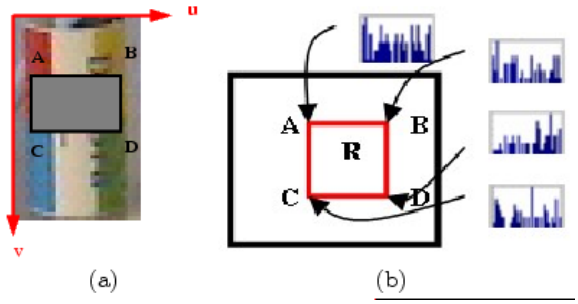


Fig. 5. Integral Images, a) Integral Image b) Orientation Integral Image

5 The Local Orientation Integral Image

An integral image is a representation of the image that allows a fast computation of features because it does not work directly with the original image intensities. Instead, it works over an incrementally built image that adds feature values along rows and columns. Once computed this image representation, any one of the local features (weak classifiers) can be computed at any location and scale in constant time.

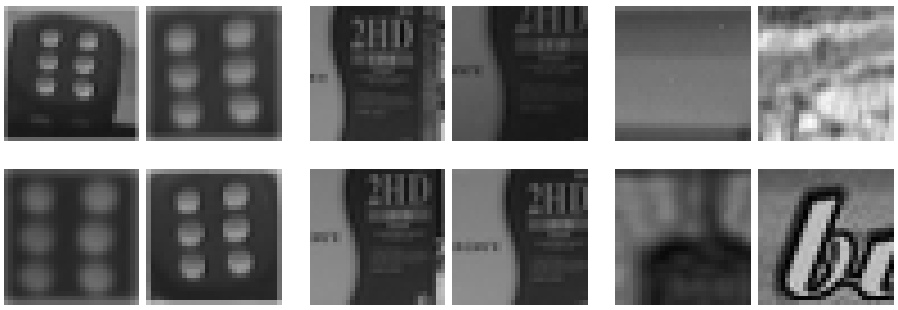
In its most simple form, the value of the integral image M at coordinates u, v contains the sum of pixels values above and to the left of u, v, inclusive.

$$M(u,v) = \sum_{i \leq u, j \leq v} I(i,j) \tag{9}$$

Then, it is possible to compute for example, the sum of intensity values in a rectangular region simply by adding and subtracting the cumulative intensities at its four corners in the integral image (Figure 5a). Then, the response from the Haar-filters can be calculated in a fast way independently of size or location.

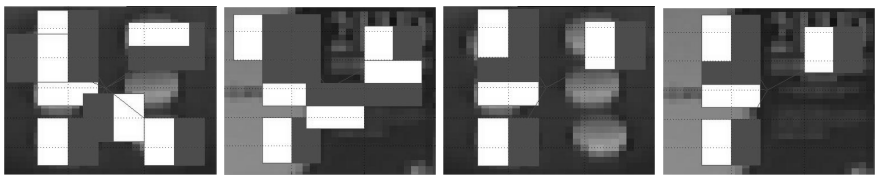
$$\text{Area} = A + D - B - C \tag{10}$$

Extending the idea of having cumulative data at each pixel in the Integral Image, we decide to store in it orientation histogram data instead of intensity



a) b) c)

Fig. 6. Training object classes. a) dice images, b) CD box images, and c) background images.



(a) (b) (c) (d)

Fig. 7. Constellations. a) dice constellation b) CD box constellation (c-d) joint classifiers

sums. Once constructed this orientation integral image, it is possible to compute a local orientation histogram for any given rectangular area within an image in constant time. see Figure 5b.

$$\begin{aligned}\text{Histogram(Area)} = \text{Histogram}(A) + \text{Histogram}(D) \\ -\text{Histogram}(B) - \text{Histogram}(C)\end{aligned} \tag{11}$$

6 Experiments

In this communication we report on initial recognition results for a limited number of objects in gray scale images. The training set had 100 images for each class, and 500 negatives or background images. These negatives images were extracted from exterior and interior scenes. The positive class images used for training presented some small translation, orientation, and scale, as shown in Figure 6.

Figure 7 a) and b) show examples of extracted feature constellation for each object class. Each one is composed by 8 weak classifiers (Haar-like features), with 4 of them common to both classes, and the remaining 4 specific to each class.

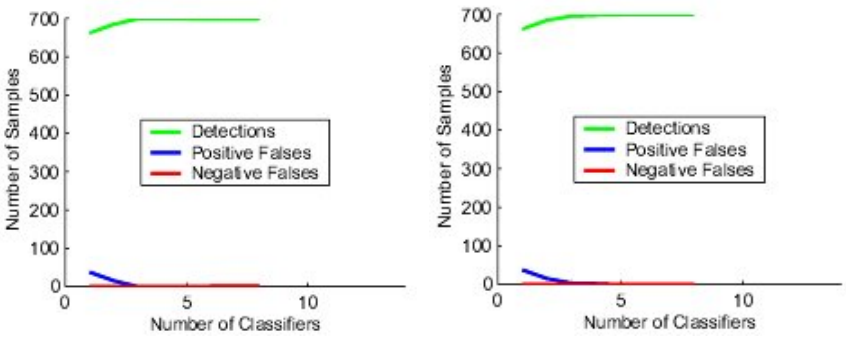


Fig. 8. Training performance. a) dice b) CD box

Thus, producing a hierarchical structure of weak classifiers. Frames c) and d) show only those four classifiers that are common to both classes. They capture simmilar local information in both classes, separating them from the background set, without the need to be class specific.

The Strong Classifiers can be expressed as the combination of joint and specific weak classifiers. Consider the dice to be class 1, the CD box to be class 2, and c_{12} the set of training samples containing either one or both objects. Then

$$H(I, c_1) = \sum h(I, c_{12}) + \sum h(I, c_1) \tag{12}$$

$$H(I, c_2) = \sum h(I, c_{12}) + \sum h(I, c_2) \tag{13}$$

The training curves are shown in Figure 8.They illustrate how the correct classification of the training set is achieved. Some results in detection process over a image sequence are visualized in Figure 9.

7 Conclusions

In this paper we have introduced a hierarchical feature selection structure that reduce the total number of weak classifiers needed to detect multiples object classes. With this method the system finds common features among objects and generalizes the detection problem.

Our approach is based on boosting over a set of simple local features. In contrast to previous approaches, and to efficiently cope with orientation changes, we propose the use of Haar basis functions and a new orientation integral image for a speedy computation of local orientation.

References

1. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2) (2004) 91–110

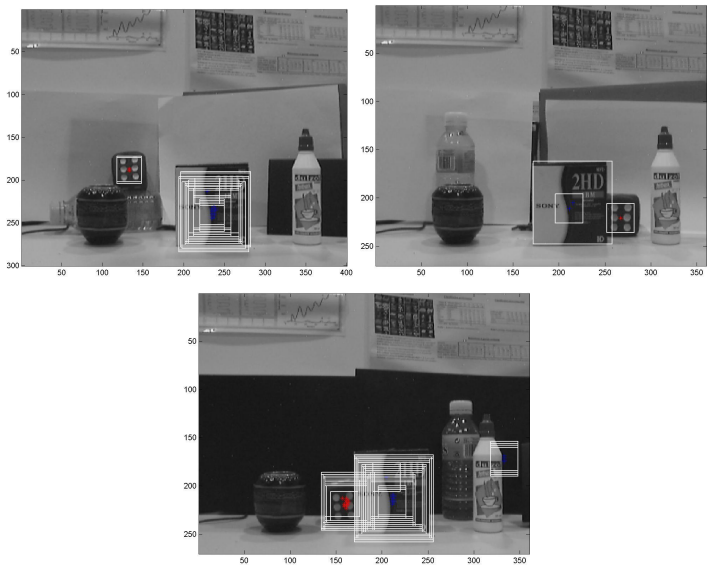


Fig. 9. Examples of correct detection of classifiers trained jointly (dice and Cd box). The last image shows also under what circumpstances a false detection might occur.

2. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. 15th IEEE Conf. Comput. Vision Pattern Recog., Kauai (2001) 511–518
3. Li, L.: Multiclass boosting with repartitioning. In: Proc. 23rd Int. Conf. Machine Learning, Pittsburgh (2006) To appear.
4. Eibl, G., Pfeiffer, K.P.: Multiclass boosting for weak classifiers. J. Mach. Learn. Res. **6** (2005) 189–210
5. Torralba, A., Murphy, K., Freeman, W.: Sharing features: efficient boosting procedures for multiclass object detection. In: Proc. 18th IEEE Conf. Comput. Vision Pattern Recog., Washington (2004) 762–769
6. Villamizar, M., Sanfeliu, A., Andrade-Cetto, J.: Computation of rotation local invariant features using the integral image for real time object detection. In: Proc. 18th IAPR Int. Conf. Pattern Recog., Hong Kong, IEEE Comp. Soc. (2006) To appear.
7. Papageorgiou, C.P., Oren, M., Poggio, T.: A general framework for object detection. In: Proc. IEEE Int. Conf. Comput. Vision, Bombay (1998) 555
8. Yokono, J., Poggio, T.: Oriented filters for object recognition: An empirical study. In: Proc. 6th IEEE Int. Conf. Automatic Face Gesture Recog., Seoul (2004) 755–760
9. Yokono, J., Poggio, T.: Rotation invariant object recognition from one training example. Technical Report 2004-010, MIT AI Lab. (2004)
10. Freeman, W.T., Adelson, E.H.: The design and use of steerable filters. IEEE Trans. Pattern Anal. Machine Intell. **13**(9) (1991) 891–906
11. Schaffalitzky, F., Zisserman, A.: Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". In: Proc. 7th European Conf. Comput. Vision, Copenhagen, Springer-Verlag (2002) 414–431