Combining Densely Sampled Form and Motion for Human Action Recognition

Konrad Schindler[1,⋆] and Luc van Gool[1,2]

[1] BIWI / ETH Zürich, Sternwartstrasse 7, CH-8092 Zürich, Switzerland
[2] ESAT / KU Leuven, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium

Abstract. We present a method for human action recognition from video, which exploits both form (local shape) and motion (local flow). Inspired by models of the human visual system, the two feature sets are processed independently in separate channels. The form channel extracts a dense local shape representation from every frame, while the motion channel extracts dense optic flow from the frame and its immediate predecessor. The same processing pipeline is applied in both channels: feature maps are pooled locally, down-sampled, and compared to a collection of learnt templates, yielding a vector of similarity scores. In a final step, the two score vectors are merged, and recognition is performed with a discriminative classifier. In an evaluation on two standard datasets our method outperforms the state-of-the-art, confirming that the combination of form and motion improves recognition.

1 Introduction

Recognising human actions in monocular video is an active area of computer vision research, with applications in diverse fields including surveillance, content-based video search, and human-computer interaction.

Possible features used to describe human actions include dense optic flow [9,16], space-time silhouettes [3,26], stick figures obtained from those silhouettes [1], or sparse space-time interest points [8,18,19].

A common property of most previous approaches is that a *single* type of feature is extracted: the underlying assumptions are either that local motion is sufficient to recognise human actions (for flow-only methods), or that feature extraction can be designed such that the features capture both the local motion and the shape/appearance of the human body.

In this work, we present a method, which independently extracts *dense shape and motion features in separate processing channels*, and only fuses them for the final classification stage. This approach is inspired by the human visual system, which is thought to extract and process shape and motion separately, in distinct regions of the visual cortex [5,10,14]. The only other work we are aware of, which

⋆ The authors acknowledge support from the EU 6th framework projects COBOL (NEST-043403) and DIRAC (IST-027787).

G. Rigoll (Ed.): DAGM 2008, LNCS 5096, pp. 122–131, 2008.

has combined shape and motion features, is [19], which extracts two sparse sets of interest points (one in image space and one in the 3D space-time volume).

As far as we know, the present work is the first practical implementation of a full "biologically inspired" system with a form as well as a motion channel. The only other system which simulates both channels is described in the seminal work of Giese and Poggio [14]. Their work has inspired ours, however they were mainly interested in the neurological plausibility of the model, and did not go beyond a proof-of-concept implementation for simple, schematic stimuli.

An experimental evaluation is presented on two standard datasets, which confirms that using independent shape and motion channels indeed boosts recognition performance. In a comparison with state-of-the-art methods, our system achieves the highest published recognition rates to date.

2 Related Work

Early approaches to human action recognition used the tracks of body parts as input features [21,27]. Since an action is defined by an articulated motion pattern, this choice is obvious. However, it depends critically on correctly tracking the parts of the human body in monocular video, a notoriously difficult task.

In [4], action recognition is cast as shape matching: an action is represented by a single unambiguous pose, and recognition reduces to comparing poses (described by edge maps). The success of this method demonstrated the importance of shape, while most later works have focused on the dynamic aspect of human actions. However, direct shape matching assumes that a sufficient part of the body contour can be found with edge detection, which is not guaranteed under realistic imaging conditions.

More recently, researchers moved away from high-level representations of the body, and instead use a collection of low-level features, which is less compact and less intuitive, but more robust. Efros et al. [9] apply optic flow filters to a window around the person, and feed the filter responses into an exemplar-based classifier. Jhuang et al. [16] also use optic flow: they extend the static recognition model [24], by replacing Gabor filters with flow filters. Filter responses are pooled locally, and filtered again with more complex flow templates learnt from examples. The new responses are pooled again and passed to a discriminative classifier. The architecture is inspired by the human visual system, and is similar in spirit to the *motion channel* of our work.

In [18], the video sequence is represented by a sparse set of spatio-temporal corners found with a 3D version of the Harris detector. Different descriptors are proposed for the interest points, based on either space-time gradients, or on optic flow. The set of interest points is classified with either nearest-neighbour matching [18], or a SVM [23]. The method of [8] is conceptually similar, but with a different spatio-temporal interest point detector based on 1D Gabor filters. Optic flow descriptors at these points are quantised to a fixed vocabulary of space-time visual words, and recognition is performed by matching visual word histograms. This method was also extended to unsupervised learning with pLSA [20].

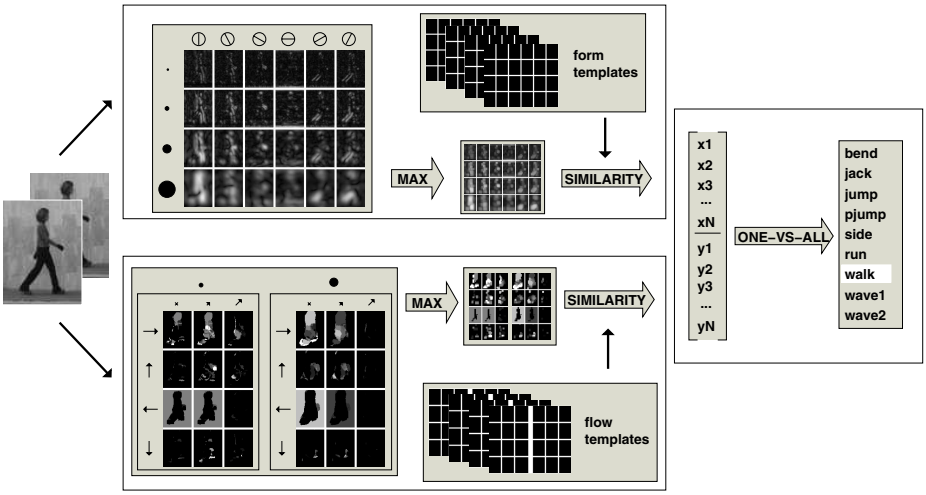


Fig. 1. Overview of the recognition system. Features are extracted in two parallel processing streams. In the form channel (top), log-Gabor filters at multiple orientations and scales are applied. The motion channel (bottom) extracts optic flow at different scales, directions, and speeds. Filter responses in both channels are MAX-pooled, and compared to a set of learnt templates. The similarity values are concatenated to a single feature vector and are classified with a bank of linear classifiers.

Niebles and Li [19] represent a frame by sparse sets of local features. The extract appearance descriptors at spatial interest points, as well as spatio-temporal descriptors at space-time interest points (using the method of [8]). A constellation model is learnt for the features and used to train a discriminative classifier. Other than the previously described works, this one combines two independently computed types of features. Our work follows the same idea, but is different in that our base features are densely sampled, and in that [19] does not make a clear distinction between form and motion features.

Blank et al. [3] represent actions by 3D shapes formed by sequences of silhouettes. Local properties of these "space-time shapes" are extracted from the Poisson equation, and used as features for exemplar-based classification. Wang and Suter [26] also start from a sequence of silhouettes, and extract features with Kernel PCA. A Conditional Random Field is trained to classify new sequences.

Ali et al. [1] return to an articulated model. Skeletonisation of silhouettes is used to obtain 2D trajectories of the main joints, and a set of chaotic invariants of the trajectories serves as features for a kNN-classifier.

3 System Details

Our system independently processes dense form (shape) and motion (flow) features, in what is sometimes called a "biologically inspired" manner due to the similarity with the ventral and dorsal pathways of the primate visual cortex [10]:

two sets of low-level cues are extracted from the data with independent algorithms and are separately converted to sets of high-level features. The idea is to minimise correlations between the two sets and in this way provide a richer description to the actual recognition system, which in our case is a discriminative classifier. Figure 1 illustrates the complete processing pipeline.

Feature extraction and action recognition are performed independently for every frame. Similar to other frame-based methods [16,19], the action labels assigned to individual frames are converted to a sequence label with a simple majority vote, corresponding to a "bag-of-frames" model.

3.1 Input Data

Like [9,16], we use a person-centred coordinate frame: our input is a sequence of fixed-size image windows, centred at the person of interest. Note that is a subtle difference to [9]: they assume that the person is seen on a uniform background, so that only the relative articulations are extracted. In contrast, our method, and also [16], can see the inverse flow of the background, and thus take into account a person's motion through the image coordinate frame.

Other than for silhouette-based methods [1,3,26], no figure-ground segmentation is required, thus making the method more widely applicable. In particular, reliable silhouette extraction in practice requires a static background. In contrast, human detectors based on sliding windows (such as [7]) and trackers based on rectangular axis-aligned regions (such as [6]) naturally provide bounding boxes.

3.2 Form Features

Local shape is extracted from each frame separately with a bank of Gabor-like filters. Gabor filtering is a standard way to find local oriented edges (similar to the simple cells of Hubel and Wiesel [15] in area V1 of the visual cortex). Specifically, we use log-Gabor filters, which allow a better coverage of the spectrum than the standard (linear) version with fewer preferred frequencies, and are also consistent with electro-physiological measurements [11]. The response g at position (x, y) and spatial frequency w is

$$g^w(x,y) = \frac{1}{\mu} \left\| e^{-\frac{\log(w(x,y)/\mu)}{2\log\sigma}} \right\| , \tag{1}$$

with μ the preferred frequency of the filter, and σ a constant, which is set to achieve even coverage of the spectrum. $\|\cdot\|$ denotes the magnitude of the (complex) response. The phase is discarded. The filter gain is adapted to the frequency spectrum of natural images: the gain factor is proportional to the frequency, to give all scales equal importance. We filter with 6 equally spaced orientations and 4 scales (see Table 1 for parameter values of the implementation).

To increase robustness to translations, each orientation map is down-sampled with the MAX-operator. Using the MAX, rather than averaging responses, was originally proposed by [12] and has been strongly advocated by [22], because of its

ability to preserve contrast features and its conformity with electro-physiological measurements [13,17]. The response at location (x, y) is given by

$$h(x,y) = \max_{(i,j)\in\mathcal{G}(x,y)} \left[\, g(i,j) \,\right] , \tag{2}$$

where $\mathcal{G}(x, y)$ denotes the receptive field (local neighbourhood) of the pixel (x, y). Our receptive field size of 9×9 pixels (determined experimentally) agrees with the findings of [16,25] (see also Table 1).

In a last step, the orientation patterns are compared to a set of templates, resulting in a vector $\mathbf{q}_f$ of similarity scores. In order to learn an informative set of templates, the pooled orientation maps are rearranged into one vector $\mathbf{h}$ per frame, and simple linear PCA is applied. A fixed number N of basis vectors $\{\mathbf{b}_i, i = 1 \ldots N\}$ are retained and are directly viewed as templates for relevant visual features.

The incoming vector $\mathbf{h}$ from a new image is scaled to norm 1 (corresponding to a normalisation of signal "energy"), and projected onto the set of templates. The linear projection $\langle \mathbf{h}, \mathbf{b}_i \rangle = \cos(\angle_{\mathbf{h}}^{\mathbf{b}_i})$ onto template $\mathbf{b}_i$ can be directly interpreted as a similarity measure, where 1 means that the two are perfectly equal, and 0 means that they are maximally dissimilar. In our implementation, we use 500 templates (see also Table 1). Note that we learn a set of templates, which is in some sense optimal to describe the training data (in the same way in which PCA is), whereas [24,16] use random features to enable a biologically more plausible learning rule.

3.3 Motion Features

In discrete video, optic flow has to be computed between neighbouring frames. By convention, we regard the optic flow computed between consecutive frames as a feature of the second frame. Hence, the "motion features at frame t" are the ones computed from the flow field between frames $(t-1)$ and t.

At every frame, dense optic flow is estimated directly, by template matching to the previous frame with the L_1-distance (sum of absolute differences). Although optic flow is notoriously noisy, we do not smoothe it in any way, for the following reasons: spatial smoothing blurs the flow field at discontinuities, where the information is most important, if no figure-ground segmentation is available; smoothing over time is incompatible with a causal video processing system, since it requires future data.

To obtain a representation analogous to the log-Gabor maps for form, the optic flow is discretised into a set of response maps for different "flow filters", each with different preferred flow direction and speed. A filter's response $r(x, y)$ is maximal, if the direction and speed at location (x, y) exactly match the preferred values, and decreases linearly with changing direction and/or speed. Responses are computed at 2 spatial scales, 4 equally spaced directions (half-wave rectified), and 3 scale-dependent speeds (see Table 1 for parameter values of our implementation).

Table 1. Summary of parameter values used in our implementation

step	parameter	form channel	motion channel	
low-level filtering	directions [°]	0, 30, 60, 90, 120, 150	0, 90, 180, 270	
	scales [pix]	2, 4, 8, 16	8	16
	velocities [pix/frame]	—	0, 2, 4	0, 4, 8
MAX-pooling	sampling step [pix]	5	5	
	window size [pix]	9	9	
high-level features	# templates	500	500	
	relative weight	0.3	0.7	

The remaining processing steps of the flow channel are the same as for the form channel. Flow maps are MAX-pooled and converted to a vector $\mathbf{q}_m$ of similarity values by comparing to a set of flow templates learnt with PCA. Note that although we compute optic flow without smoothing, the templates are smooth due to the denoising effect of PCA. The same parameters are used in the form and flow pathways (see Table 1).

3.4 Classifier

For classification, the feature vectors for form and motion are simply concatenated, $\mathbf{q} = [(1-\lambda)\mathbf{q}_f\ ,\ \lambda\mathbf{q}_m]$. The weight $\lambda \in [0..1]$ has been determined experimentally, and has proven to be stable across different datasets, see Section 4.1.

To classify an action into one of K classes, we train a bank of linear *one-vs-all* SVMs, each with weights $W = (K-1)$ for in-class samples, and $W = 1$ for out-of-class samples (to account for the uneven number of samples). We purposely keep the classifier simple. Using non-linear kernels for the SVM yields no significant improvement – presumably due to the high dimensionality of the data. Alternative multi-class extensions, such as the *all-pairs* strategy, give similar results in practice, but require a larger number of binary classifiers. Note also that *one-vs-all* has an obvious interpretation in terms of biological vision, with each binary classifier representing a neural unit, which is activated only by a certain action class.

4 Results

We use two standard datasets for our evaluation (examples are shown in Figure 2), which in the last few years have become the de-facto standards for benchmarking human action recognition.

The WEIZMANN set was originally recorded for [3], and consists of 9 subjects performing a set of 9 different actions: *bending down, jumping jack, jumping, jumping in place, galloping sideways, running, walking, waving one hand, waving both hands.* Bending down is only shown once, all other action are periodic, and are repeated several times. The videos have varying length. To avoid biases in the evaluation, we have trimmed all sequences to 28 frames (the length of the shortest sequence). The height of humans is in the range $\approx$62..75 pixels, and the size of

Fig. 2. Examples of actions from databases WEIZMANN (top) and KTH (bottom)

the extended bounding box (see section 3.1) is 67×89 pixels. All evaluations were done with 9-fold cross-validation: 8 subjects are used for training, the remaining 1 for testing; results are averaged over all 9 permutations.

The KTH set was recorded for [18,23], and consists of 25 subjects performing 6 different periodic actions: *boxing, hand-clapping, jogging, running, walking, hand-waving.* The height of humans is in the range ≈55..103 pixels, the size of the extended bounding box is 87×103. The complete set was recorded four times under different conditions: outdoors (S1), outdoors with scale variations (S2), outdoors with different clothes (S3), and indoors (S4). *Jogging*, *running*, and *walking* are performed multiple times in each video, separated by large numbers of empty frames (which obviously cannot be classified). We therefore use only one pass. All sequences are trimmed to the minimum length of 18 frames. Evaluations were done with 5-fold cross-validation: the data is split into 5 folds of 5 subjects each; in turn, 4 folds are used for training, and the remaining 1 for testing.

To account for the symmetry of the human body and its motion, we also mirror all sequences along the vertical axis, for both training and testing. The same parameter setting were used for all reported experiments (see Table 1).

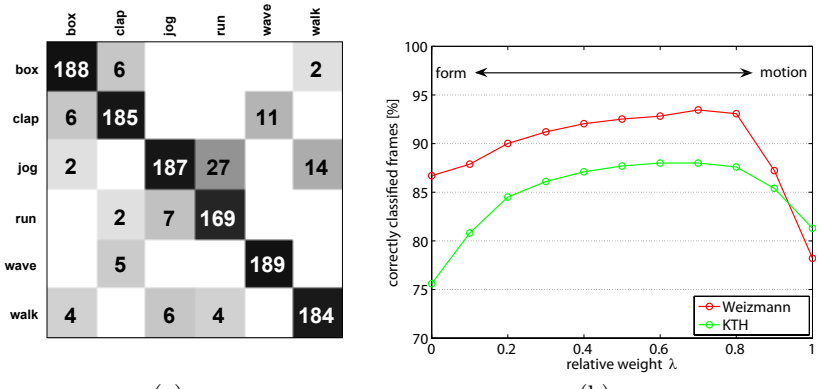
A comparison of recognition results is given in Table 2. We begin with the WEIZMANN dataset, which is currently the most widely used benchmark for action recognition. Our method achieves perfect recognition. As a side-note, we believe the dataset is too easy, and the community should adopt a more challenging one as a standard (even the first publication for which it was created reported perfect results [3]). On the KTH dataset, we achieve the best result to date – see Table 2. The full confusion matrix for the 1198 videos is given in Figure 3(a). Note that confusions occur mostly between the pairs *run–jog* and *walk–jog*, which intuitively makes sense. Especially the semantic boundary between running and jogging is not well-defined. The comparison should be taken with a grain of salt: in the action recognition literature there is no established testing protocol, and the sizes of training and test sets vary between different works. We always quote the best results someone has published. Still, the comparison remains indicative.

Table 2. Comparison of recognition results (percentage of correctly classified action videos). On both datasets, our method achieves the highest performance. In JHUANG the four scenarios in the KTH dataset were processed separately, as independent datasets with smaller variation.

KTH		WEIZMANN	
this paper	**92.7** %	this paper	**100.0** %
NIEBLES [20]	81.5 %	BLANK [3]	**100.0** %
DOLLÁR [8]	81.2 %	JHUANG [16]	98.8 %
SCHÜLDT [23]	71.7 %	WANG [26]	97.8 %
JHUANG [16]	91.7 %	ALI [1]	92.6 %
(average of scenarios s1–s4, trained and tested separately)		DOLLÁR [16]	86.7 %
		NIEBLES [19]	72.8 %

4.1 Contributions of Form and Motion Features

One of the aims of the present work is to exploit both form *and* motion as cues for action recognition in videos. It is obvious to ask, what benefit this brings, and how to combine the two channels. We have run experiments with our system, in which we have changed the relative weight λ of the two cues – including the cases where one channel is not used at all. In all our experiments on the two datasets, as well as on smaller subsets, a combination of form and motion outperforms both form alone and motion alone. Furthermore, it turns out that the optimal relative weight is approximately the same across different data sets, and the exact choice within a reasonable range is not critical (for our system $\lambda = 0.7$,



(a) (b)

Fig. 3. (a) Confusion matrix for the KTH dataset. The database contains 1198 videos, 1102 of which are classified correctly. Most confusions happen between the visually similar classes *run–jog–walk*. (b) Effect of using a form and a motion channel. The combination of the two consistently outperforms form alone ($\lambda = 0$) as well as motion alone ($\lambda = 1$). The best mixing coefficient is the same for both datasets. Note that the y-axis is the percentage of correctly recognised *frames*, not sequences.

but being a normalisation between two independently computed feature sets, the value may depend on both the specific implementation of feature extraction, and the parameter settings).

Results are shown in Figure 3. The evaluation is done with the classification results for single frames, not entire sequences, since the majority vote used to determine the sequence label is not a smooth mapping.

It is interesting to note that for the WEIZMANN dataset form alone is a better cue than flow alone, while for the KTH set it is the other way round. This suggests that our implementation does not introduce a strong bias towards one or the other feature, and supports the claim that explicit channels for both cues improves recognition performance.

5 Concluding Remarks

We have presented a method for human action recognition, which is based on independently computed form and motion features, sampled densely from the image plane. An experimental evaluation of the method has confirmed the advantage of explicitly extracting both form and motion. The method performs well on different databases without any parameter changes. Training the system is computationally expensive, the bottleneck being the singular value decomposition for template learning. On the contrary, testing a new image is efficient. In our un-optimized implementation it takes a few seconds, but the bottleneck is the optic flow computation, which can be done at frame-rate on modern GPUs [28].

A limitation of the system is that it currently does not have mechanisms for invariance against scale, rotation, and viewpoint changes (although in the KTH database, it successfully handles scaling up to a factor of ≈ 2, and viewpoint changes up to $\approx 30^\circ$).

An open questions, which remains to be investigated, is which frames within a motion sequence reliably support its classification, and which frames are ambiguous. This is directly related to a classic question of both computer vision [4] and neuro-science [2], namely which frames can serve as "key-frames" and *have to be* seen to recognise the action.

References

1. Ali, S., Basharat, A., Shah, M.: Chaotic invariants for human action recognition. In: Proc. ICCV (2007)
2. Beintema, J.A., Lappe, M.: Perception of biological motion without local image motion. P. Natl. Acad. Sci. USA 99, 5661–5663 (2002)
3. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: Proc. ICCV (2005)
4. Carlsson, S., Sullivan, J.: Action recognition by shape matching to key frames. In: Proc. Workshop on Models versus Exemplars in Computer Vision (2001)
5. Casile, A., Giese, M.A.: Critical features for the recognition of biological motion. J. Vision 5, 348–360 (2005)

6. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE T. Pattern Anal. 25(5), 564–575 (2003)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc ICCV, pp. 886–893 (2005)
8. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: Workshop on Performance Evaluation of Tracking and Surveillance (VS-PETS) (2005)
9. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: Proc. ICCV (2003)
10. Felleman, D.J., van Essen, D.C.: Distributed hierarchical processing in the primate visual cortex. Cereb. Cortex 1, 1–47 (1991)
11. Field, D.J.: Relations between the statistics of natural images and the response properties of cortical cells. J. Opt. Soc. Am. A. 4(12), 2379–2394 (1987)
12. Fukushima, K.: Neocognitron: a self-organizing neural network model for mechanisms of pattern recognition unaffected by shift in position. Biol. Cybern. 36, 193–202 (1980)
13. Gawne, T.J., Martin, J.: Response of primate visual cortical V4 neurons to two simultaneously presented stimuli. J. Neurophysiol. 88, 1128–1135 (2002)
14. Giese, M.A., Poggio, T.: Neural mechanisms for the recognition of biological movements. Nat. Neurosci. 4, 179–192 (2003)
15. Hubel, D.H., Wiesel, T.N.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. (Lond) 160, 106–154 (1962)
16. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: Proc. ICCV (2007)
17. Lampl, I., Ferster, D., Poggio, T., Riesenhuber, M.: Intracellular measurements of spatial integration and the max operation in complex cells of the cat primary visual cortex. J. Neurophysiol. 92, 2704–2713 (2004)
18. Laptev, I., Lindeberg, T.: Local descriptors for spatio-temporal recognition. In: Proc. ICCV (2003)
19. Niebles, J.C., Fei-Fei, L.: A hierarchical model of shape and appearance for human action classification. In: Proc. CVPR (2007)
20. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatio-temporal words. In: Proc. BMVC (2006)
21. Rao, C., Yilmaz, A., Shah, M.: View-invariant representation and recognition of actions. Int. J. Comput. Vision 50(2), 203–226 (2002)
22. Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nat. Neurosci. 2, 1019–1025 (1999)
23. Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: Proc. ICPR (2004)
24. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Object recognition with cortex-like mechanisms. IEEE T. Pattern Anal. 29(3), 411–426 (2007)
25. Serre, T., Wolf, L., Poggio, T.: Object recognition with features inspired by visual cortex. In: Proc. CVPR (2005)
26. Wang, L., Suter, D.: Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical model. In: Proc. CVPR (2007)
27. Yacoob, Y., Black, M.J.: Parameterized modeling and recognition of activities. Comput. Vis. Image Und. 72(2), 232–247 (1999)
28. Zach, C., Pock, T., Bischof, H.: A duality-based approach to realtime $TV-L_1$ optical flow. In: Hamprecht, F.A., Schnörr, C., Jähne, B. (eds.) DAGM 2007. LNCS, vol. 4713. Springer, Heidelberg (2007)