Fast categorisation of articulated human motion

Konrad Schindler[1] **and Luc van Gool**[2]
[1] *Computer Science Dept., TU Darmstadt, Germany*
[2] *Computer Vision Laboratory, ETH Zürich, Switzerland*
and *ESAT/PSI-IBBT, K. U. Leuven, Belgium*

Abstract

Visual categorisation of human motion in video clips has been an active field of research in recent years. However, most published methods either analyse an entire video and assign it a single category label, or use relatively large look-ahead to classify each frame. Contrary to these strategies, the human visual system proves that simple categories can be recognised almost instantaneously. Here we present a system for categorisation from very short sequences ("snippets") of 1–10 frames, and systematically evaluate it on several data sets. It turns out that even local shape and optic flow for a single frame are enough to achieve $\approx$ 80-90% correct classification, and snippets of 5-7 frames (0.2-0.3 seconds of video) yield results on par with the ones state-of-the-art methods obtain on entire video sequences.

1 INTRODUCTION

Recognising human motion categories in monocular video is an important scene understanding capability, with applications in diverse fields such as surveillance, content-based video search, and human-computer interaction. By *motion categories* we mean a semantic interpretation of the articulated human motion. Most computer vision research in this context has concentrated on human *action* recognition, while we only see actions as one possible set of semantic categories, which can be inferred from the visual motion pattern. We will also show a more subtle example, in which emotional states are derived from body language.

Past research in this domain can be roughly classified into two approaches: one that extracts a *global* feature set from a video (Ali et al., 2007; Dollár et al., 2005; Laptev and Lindeberg, 2003; Wang and Suter, 2007), and using these features aims to assign a single label to the *entire video* (typically several seconds in length). This paradigm obviously requires that the observed motion category does not change during the duration of the video.

The other approach extracts a feature set *locally* for a frame (or a small set of frames), and assigns an *individual* label to each frame (Blank et al., 2005; Efros et al., 2003; Jhuang et al., 2007; Niebles and Fei-Fei, 2007). If required, a global label for the sequence is obtained by simple voting mechanisms. Usually these methods are not strictly causal: features are computed by analysing a temporal window centred at the current frame, therefore the classification lags behind the observation—to classify a frame,future information within the temporal window is required.

Both approaches have achieved remarkable results, but human recognition performance suggests that they might be using more information than necessary: we can correctly recognise motion patterns from very short sequences—often even from single frames.

1.1 Aim of this work

The question we seek to answer is *how many frames are required to categorise motion patterns?* As far as we know, this is an unresolved issue, which has not yet been systematically investigated (in fact, there is a related discussion in the cognitive sciences, see section 4). However, its answer has wide-ranging implications. Therefore, our goal is to establish a baseline, how long we need to observe a basic motion, such as *walking* or *jumping*, before we can categorise it.

We will operate not on entire video sequences, but on very short sub-sequences, which we call *snippets*. In the extreme case a snippet can have length 1 frame, but we will also look at snippets of up to 10 frames. Note that in many cases a single frame is sufficient, as can be easily verified by looking at the images in Fig. 1. The main message of our study is that *very short snippets (1-7 frames), are sufficient to distinguish a small set of motion categories, with rapidly diminishing returns, as more frames are added.*

Figure 1: *Examples from databases* WEIZMANN *(left),* KTH *(middle), and* LPPA *(right). Note that even a single frame is often sufficient to recognise an action, respectively emotional state.*

This finding has important implications for practical scenarios, where decisions have to be taken online. Short snippets greatly alleviate the problem of temporal segmentation: if a person switches from one action to another, sequences containing the transition are potentially problematic, because they violate the assumption that a single label can be applied. When using short snippets, only few such sequences exist. Furthermore, short snippets enable fast processing and rapid attention switching, in order to deal with further subjects or additional visual tasks, before they become obsolete.

1.2 Method

To investigate the influence of snippet length on the categorisation performance, we present a *causal* categorisation method, which uses only information from few past frames. The method densely extracts form (local edges) and motion (optic flow) from a snippet, and

separately compares them to learnt templates. The similarity scores for both feature sets are concatenated to a single vector and passed on to a classifier. As far as we know, this is also the first practical implementation of a "biologically inspired" system with both a form and a motion pathway. The strategy to process form and motion independently with a similar sequence of operations was inspired by the seminal work of Giese and Poggio (2003). In their paper, they describe the only other simulation of both pathways, however their proof-of-concept implementation is only designed for simple, schematic stimuli.

In detailed experiments we evaluate the effect of changing the snippet length, as well as the influence of form and motion features. We also compare to other methods on two standard data sets, both at the level of snippets and whole sequences, and obtain results on par with or better than the state-of-the-art.

2 PRELIMINARIES

A basic assumption underlying most of the literature on automatic "action recognition" is that perception is (at least to a certain degree) *categorical*, meaning that an observed pattern can be assigned to one out of a relatively small set of categories. This assumption is in line with the prevalent view in the cognitive sciences (e.g. Harnad, 1987). The categorical approach has several advantages: in terms of scene understanding, categories offer a discrete, higher-level abstraction of the continuous space of articulated motions; on the computational side, categorisation is a discriminative task, and can be solved without a complete generative model of human behaviour.

2.1 Defining categories

In the previous section, we have used the term *basic* motion to describe the units, into which the articulated motion shall be classified. The reason for this terminology is that there is another unresolved issue looming behind our question, namely the definition of what constitutes a motion category. Obviously, the amount of information which needs to be accumulated, and also the number of relevant classes for a given application, both depend on the complexity (e.g., recognising a high-jump takes longer than separately recognising the three components *running*, *jumping*, and *falling on the back*). This leads to the problem of decomposing motion patterns: can, and should, complex motion patterns be decomposed into sequences of simpler "atomic motions", which again can be recognised quickly?

The decomposition problem appears to be application-dependent, and is *not* the topic of this study. We assume that a relatively small set of basic categories, such as *walking* or *waving*, form the set of possible labels, and that the labels are relatively unambiguous.[1] These assumptions have been made implicitly in most of the published work on action recognition, as can be seen from the standard databases (the ones also used in this work).

[1] In practice, some categories will by their nature be ambiguous and lack well-defined boundaries, e.g. *jogging* vs. *running* , or *sad walking* vs. *neutral walking* performed by a tired person.

2.2 Motion Snippets

The aim of the present work is not only to introduce yet another motion classification method, but also to systematically investigate, how much information needs to be accumulated over time to enable motion classification. In a setup with discrete time steps, this boils down to the question, how many frames are required.

Very short snippets provide less data to base a decision on, hence it becomes important to extract as much information as possible. We will therefore collect both shape information from every frame and optic flow. In real video with discrete time steps, optic flow has to be computed between neighbouring frames. By convention, we will regard the optic flow computed between consecutive frames $(t-1)$ and t as a feature of frame t. Hence, when we refer to a *snippet of length 1*, or a *single frame*, this flow field is included. In the same way, a snippet of length, say, $L=7$ comprises the shape descriptors for 7 frames, and 7 flow fields (not 6).

As will be demonstrated in section 5, using both shape and flow yields a marked improvement in categorisation performance, compared to shape alone, or flow alone.

2.3 The "right" snippets

It is obvious that not all parts of a video carry the same amount of category information. Some parts of a motion (or motion cycle) are highly distinctive, while others may be totally ambiguous. This leads to the question which "key-snippets" of a motion sequence are most suitable, respectively unsuitable, to determine a given class. It is trivial, but sometimes overlooked, that the best key-frames or key-snippets are not a fixed property of a category's motion pattern, but depend on the entire set of categories in a given application (the most suitable snippets being the ones which are rarely confused with any of the other categories).

A framework based on discriminative classification provides an easy way to identify particularly distinctive (or ambiguous) snippets, by looking at their classification margins. In the experiments section, we will show some examples of this analysis. Note that, although not investigated further in the present work, one could use this information as a confidence measure in a recognition system with variable snippet length. Such a system would then be able to decide online, whether it can reliably classify a snippet, or needs to accumulate more frames.

3 RELATED WORK

Early attempts at human action recognition used the tracks of a person's body parts as input features (Fanti et al., 2005; Rao et al., 2002; Yacoob and Black, 1999). This representation is an obvious choice, because physically the dynamics of the body parts relative to each other is what defines a motion pattern. However, it depends on correct tracking of either an articulated human model, or many separate regions, both difficult tasks, especially in monocular video.

Carlsson and Sullivan (2001) cast action recognition as a shape matching problem. An action is represented by a single unique pose, and categorisation is performed by comparing poses, described by edge maps. This demonstrated the importance of shape, while later

research focused on the dynamic aspect of human actions. In this work we will use both pieces of information.

A drawback of early approaches was that tracking, as well as contour detection, become unreliable under realistic imaging conditions. Following a general trend in computer vision, researchers therefore moved away from the high-level representation of the human body, and replaced it by a collection of low-level features, which are less compact and less intuitive, but can be extracted more reliably. Efros et al. (2003) apply optic flow filters to a window centred at the human, and use the filter responses as input to an exemplar-based classifier. Their method is probably the first one to aim for classification at the frame level from flow alone; however, although they individually label each frame, a large temporal window (up to 25 past and 25 future frames) is employed to estimate its flow.

Jhuang et al. (2007) have extended the static scene categorisation model of Serre et al. (2007), by replacing form features with motion features. Like Efros et al. (2003), they extract dense local motion information with a set of flow filters. The responses are pooled locally, and converted to higher-level responses by comparing to more complex templates learnt from examples. These are pooled again, and fed into a discriminative classifier. This approach is the most similar in spirit to our work.

Niebles and Fei-Fei (2007) also classify at the frame level. They represent a frame by sparse sets of local appearance descriptors extracted at spatial interest points, and a similar set of local motion descriptors extracted from a sub-sequence centred at the current frame, with the method of Dollár et al. (2005). A constellation model for the features is learnt, and used to train a discriminative classifier.

Laptev and Lindeberg (2003) represent an entire video sequence as a sparse set of spatio-temporal interest points, which are found with a 3D version of the Harris corner detector. Different descriptors are proposed for the space-time window around an interest point: histograms of gradients, histograms of optic flow, PCA projection of gradients, or PCA projection of optic flow. Classification is done at sequence level, either by nearest-neighbour matching (Laptev and Lindeberg, 2003), or with a SVM (Schüldt et al., 2004).

Dollár et al. (2005) present a different spatio-temporal interest point detector based on 1D Gabor filters, essentially searching for regions where the intensity changes suddenly or periodically over time. Optic flow is computed as descriptor for each 3D interest region. The set of descriptors is quantised to a fixed set of 3D visual words, and a new sequence is classified by nearest-neighbour matching of its histogram of visual words. The method was extended to unsupervised learning with pLSA by (Niebles et al., 2006).

Blank et al. (2005) extract the human silhouette from each frame, and represent the sequence as a set of "space-time shapes" defined by (overlapping) 10-frame sequences of silhouettes. Local properties of such a 3D shape are extracted from the solution of its Poisson equation, and classified with an exemplar-based nearest-neighbour classifier.

Wang and Suter (2007) also use silhouettes to classify at the sequence level. They extract features from the sequence of silhouettes by non-linear dimensionality reduction with Kernel PCA, and train a Factorial Conditional Random Field to classify new sequences.

Ali et al. (2007) return to an articulated model, but follow only the main joints to make tracking more robust. Skeletonisation is applied to silhouettes to obtain 2D stick figures, and their main joints are connected to trajectories. A video is represented by a set of chaotic invariants of these trajectories, and classified with a kNN-classifier.

4 SYSTEM DETAILS

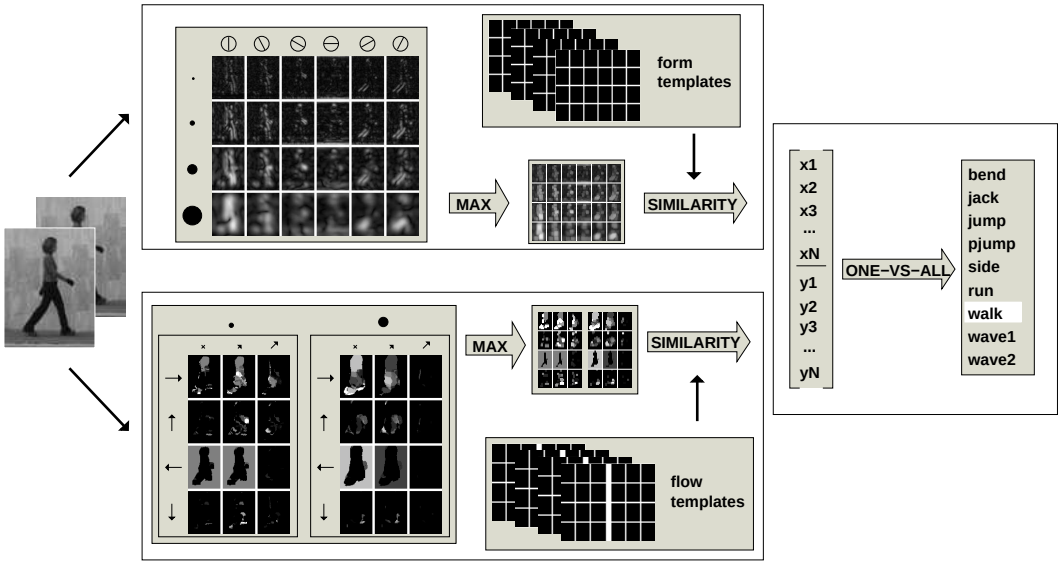

Figure 2: *System overview. From a snippet, features are extracted in two parallel processing streams. The form pathway (top) extracts local shape at multiple scales. The motion pathway (bottom) extracts local optic flow at multiple scales. In both pathways, the filter responses are* MAX*-pooled, and compared to a set of learnt templates. The similarity scores are concatenated to a feature vector, and classified with a bank of binary classifiers.*

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 (Felleman and van Essen, 1991): 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 labelling algorithm (in our case is a discriminative classifier). Fig. 2 illustrates the complete processing pipeline.

Using both types of features for motion perception is in line with the predominant view in neuro-science, (e.g. Casile and Giese, 2005), but some researchers are of the opinion, that only form information from a number of consecutive frames is required (e.g. Beintema and Lappe, 2002). The key-frame paradigm has also been explored in machine vision (Carlsson and Sullivan, 2001). Our experiments support the first view: using form and motion consistently improves categorisation, at least with our system architecture.

Feature extraction and categorisation are performed independently for every snippet. Similar to other frame-based methods (Jhuang et al., 2007; Niebles and Fei-Fei, 2007), the labels assigned to individual snippets are converted to a sequence label with a simple majority vote, corresponding to a "bag-of-snippets" model.

4.1 Input Data

We use a simple attention model to obtain a person-centred coordinate frame: our input is a sequence of fixed-size image windows, centred at the person of interest. Note that there is a subtle difference to Efros et al. (2003): 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 the one of Jhuang et al. (2007), can see the inverse flow of the background. This means that they can take into account a person's motion through the image coordinate frame, if there is enough background structure (for uniform background all three methods will behave the same, because the latter two will not be able to learn any coherent optic flow on the background).

Other than for silhouette-based methods (Ali et al., 2007; Blank et al., 2005; Wang and Suter, 2007), 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 (e.g. Dalal and Triggs, 2005) and trackers based on rectangular axis-aligned regions (e.g. Comaniciu et al., 2003) naturally provide bounding boxes.

4.2 Form Features

Local shape is extracted from each frame in a snippet 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 (1962) in area V1 of the visual cortex)—for details please refer to standard texts (e.g. Forsyth and Ponce, 2003). 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 (Field, 1987). The response g at position (x, y) and spatial frequency w is

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

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, the response map for each orientation/scale pair is down-sampled with the MAX-operator. Using the MAX, rather than averaging responses, was originally proposed by Fukushima (1980) and has been strongly advocated by Riesenhuber and Poggio (1999), because it does not blur contrast features. The MAX is also consistent with electro-physiological measurements (Gawne and Martin, 2002; Lampl et al., 2004). The response at location (x, y) is given by

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

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 Jhuang et al. (2007); Serre et al. (2005)—see also Table 1.

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

The incoming vector $\mathbf{h}^{\mathsf{F}}$ 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

$$q_i^{\mathsf{F}} = \langle \tilde{\mathbf{h}}^{\mathsf{F}}, \mathbf{b}_i^{\mathsf{F}} \rangle = \cos(\angle_{\mathbf{b}_i^{\mathsf{F}}}^{\tilde{\mathbf{h}}^{\mathsf{F}}}) \quad , \quad \tilde{\mathbf{h}}^{\mathsf{F}} = \frac{\mathbf{h}^{\mathsf{F}}}{\|\mathbf{h}^{\mathsf{F}}\|} \tag{3}$$

onto template $\mathbf{b}_i^{\mathsf{F}}$ 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 with PCA we learn a template set which is in some sense optimal (the best linear subspace to represent the training data), whereas Serre et al. (2007); Jhuang et al. (2007) use random features to enable a biologically more plausible learning rule.

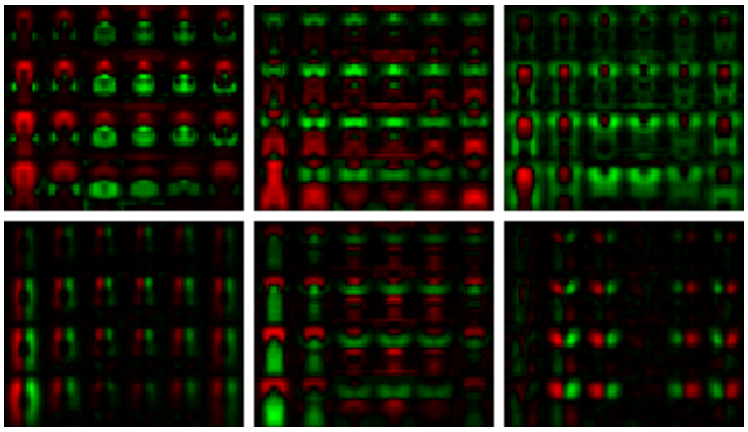

Figure 3: *Learning complex templates with PCA. Shown are the first 6 templates (basis vectors) of the* form *pathway for one run of the* WEIZMANN *dataset. Note how the templates capture complex shapes such as different arm configurations.*

4.3 Motion Features

At every frame, dense optic flow is estimated directly, by template matching: at every image location, we take the intensity pattern in a local window (receptive field), and find the most similar location in the previous frame, using the L_1-distance (sum of absolute differences). Although optic flow is notoriously noisy, we do not smooth it in any way, because smoothing blurs the flow field at discontinuities (edges in the spatial domain, direction changes in

the time domain), where the information is most important, if no figure-ground segmentation is available.

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 (receptive field sizes), 4 equally spaced directions (half-wave rectified), and 3 scale-dependent speeds (see Table 1 for parameter values of our implementation).

The remaining processing steps of the flow channel are the same as for the form channel. Flow maps are MAX-pooled to coarser maps

$$h^{\mathsf{M}}(x, y) = \max_{(i,j)\in\mathcal{G}(x,y)} [\, r(i, j) \,] \,, \tag{4}$$

and these are converted to a vector $\mathbf{q}^{\mathsf{M}}$ of similarity values by comparing to a set of flow templates learnt with PCA,

$$q_i^{\mathsf{M}} = \langle \tilde{\mathbf{h}}^{\mathsf{M}}, \mathbf{b}_i^{\mathsf{M}} \rangle \,. \tag{5}$$

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). An illustrative example of the feature extraction process is shown in Fig. 4.

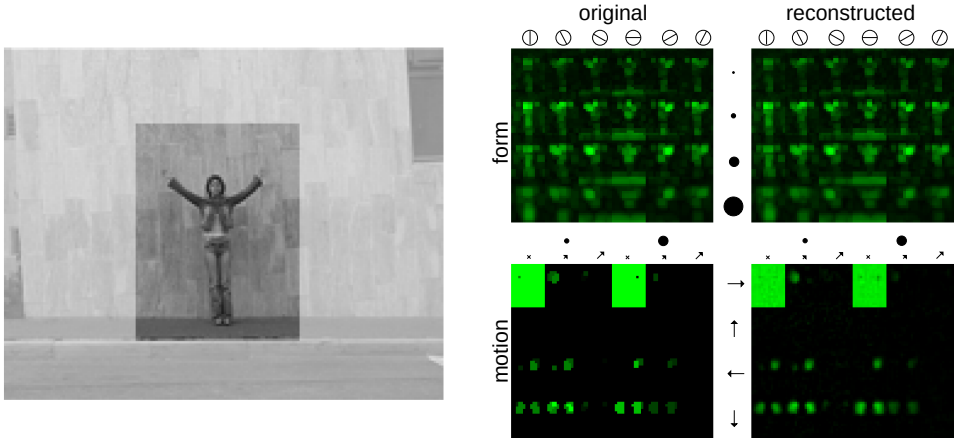

Figure 4: *Illustration of feature extraction.* Left: *image with extracted bounding box.* Middle: *feature maps after* MAX*-pooling.* Right: *feature maps reconstructed from 500 templates per channel.*

4.4 Classifier

For classification, the feature vectors for form and motion are simply concatenated, $\mathbf{q} = [(1-\lambda)\mathbf{q}^{\mathsf{F}}\,,\ \lambda\mathbf{q}^{\mathsf{M}}]$. The optimal weight $\lambda \in [0..1]$ has been determined experimentally, and has proved to be stable across different datasets, see Section 5.1.

As a classifier, we use a bank of linear *support vector machines* (SVMs). A description of the SVM algorithm is beyond the scope of this chapter, we refer the interested reader to standard texts (e.g. Shawe-Taylor and Christianini, 2000). To assign a motion to one of K classes, we train a bank of binary *one-vs-all* SVMs, i.e. each binary classifier is trained to separate one class from all others, and the class with the highest confidence (largest classification margin) is selected. During training of the binary classifiers, in-class samples are assigned weights $W = (K-1)$, and out-of-class samples are assigned weights $W = 1$, 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 category.

step	**parameter**	**form**	**motion**	
low-level filtering	direction [°]	0, 30, ..., 150	0, 90, 180, 270	
	scale [pix]	2, 4, 8, 16	8	16
	velocity [px/frame]	—	0, 2, 4	0, 4, 8
MAX-pooling	sampling step [px]	5	5	
	window size [px]	9	9	
high-level features	# templates	500	500	
	relative weight	0.3	0.7	

Table 1: *Summary of parameter values used in our implementation.*

4.5 Relation to existing methods

Since recent methods for motion categorisation, including the present one, are quite related, this section analyses some important similarities and differences.

In terms of the required preprocessing, our method uses a very crude attention model, namely a fixed-size bounding box centred at the person, like Efros et al. (2003); Jhuang et al. (2007). These three methods are less demanding than Ali et al. (2007); Blank et al. (2005); Wang and Suter (2007), which require a segmented silhouette, but more demanding than interest-point based methods (Laptev and Lindeberg, 2003; Niebles and Fei-Fei, 2007), which at least conceptually operate on the whole image—although in practice the amount of clutter must be limited, to ensure a significant number of interest points on the person. No method has yet been tested in cluttered environments with many distractors.

In terms of features used, Efros et al. (2003); Jhuang et al. (2007) extract only optic flow, Blank et al. (2005); Wang and Suter (2007) only silhouette shape. Laptev and Lindeberg (2003); Dollár et al. (2005) extract feature points in 3D space-time, and use either only flow, or combined space-time gradients as descriptors, while our work, as well as Niebles and Fei-Fei (2007), extract both cues independently.

More generally, our method belongs to a school, which favours densely sampled features over sparse interest points for visual classification problems (e.g. Dalal and Triggs,

2005; Pontil and Verri, 1998). It can also be considered a biologically inspired model, with parallels to the "standard model" of the visual cortex (Riesenhuber and Poggio, 1999): a layer of simple neurons sensitive to local orientation and local flow; a layer of pooling neurons with larger receptive fields to increase invariance and reduce the amount of data; a layer of neurons, each comparing the incoming signal to a learnt complex pattern; and a small layer of category-sensitive neurons, each firing when presented with features of a certain motion category.

The other model, which intentionally follows a biologically inspired architecture (Jhuang et al., 2007), currently only implements the motion pathway. Other than our model, it uses some temporal look-ahead to compute motion features. Furthermore, its complex templates are smaller (ours have the same size as the bounding box), and are found by random sampling, while we apply PCA to find a template set, which is in some sense optimal (albeit with a less biologically plausible learning rule).

5 EXPERIMENTAL EVALUATION

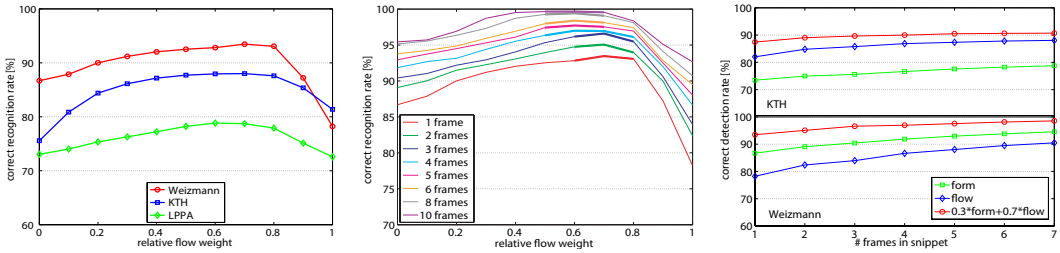

Figure 5: *Influence of form and motion features on categorisation performance.* Left: *Recognition rates at snippet length $L = 1$ with different relative weights of both pathways, computed on all three data bases.* Middle: *Recognition rates with different relative weights and different snippet lengths, shown for the* WEIZMANN *database (peaks marked by bold lines).* Right: *Recognition rates for* WEIZMANN *and* KTH *(average of all scenarios) using only form, only motion, and the best combination.*

We use three data sets for our evaluation. The first two have become the de-facto standards for human *action* recognition, while the third one has been designed for the study of *emotional body language*, i.e. emotional categories expressed in human motion.

The WEIZMANN database (Blank et al., 2005) consists of 9 subjects (3 female, 6 male) 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.* To avoid evaluation biases due to varying sequence length, we trim all sequences to 28 frames (the length of the shortest sequence). Due to the periodic nature of the actions, this gives sufficient training data, and makes sure all actions have the same influence on overall results. All evaluations on this data set were done with leave-one-out cross-validation: 8 subjects are used for training, the remaining one for testing; the procedure is repeated for all 9 permutations, and the results are averaged.

The KTH database (Laptev and Lindeberg, 2003; Schüldt et al., 2004) consists of 25 subjects (6 female, 19 male) performing 6 different actions: *boxing, hand-clapping, jog-*

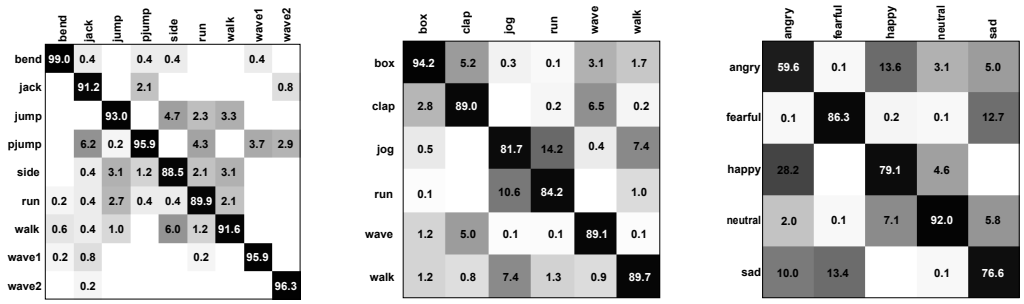

Figure 6: *Confusion matrices for $L = 1$ (form and motion at a single frame). Left to right are the true categories, top to bottom the estimated categories. Results are given in percent of the true answer—columns add up to 100.*

ging, running, walking, hand-waving. The complete set of actions was recorded under 4 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 frames, where the subject is outside the field of view. These parts are obviously meaningless in an evaluation at snippet level, so we use only one pass of each action. Again, all sequences are trimmed to the length of the shortest, in this case 18 frames. All evaluations were done with 5-fold cross-validation: the data is split into 5 folds of 5 subjects each, 4 folds are used for training, 1 for testing. The results are averaged over the 5 permutations. In the literature, KTH has been treated either as one large set with strong intra-subject variations, or as four independent scenarios, which are trained and tested separately (i.e., four visually dissimilar databases, which share the same classes). We run both alternatives.

The LPPA database has been collected for research into emotional body language. It consists of 9 trained actors (5 female, 4 male) walking in 5 different emotional styles: *neutral, happy, sad, fearful, angry.* For each subject there are 4 independent trials of each style. All sequences are trimmed to 40 frames, and evaluations are performed with leave-one-out cross-validation. This dataset is considerably harder than the schematic action datasets, because of the more subtle visual differences. As a baseline, the data was also validated in a psychophysical experiment with human observers: the full-length videos were shown to 12 subjects (5 female, 7 male), who had to assign emotion labels to them. The classification by humans was 85% correct (average over all classes).

To account for the symmetry of human body motion w.r.t. the sagittal plane, we also use all sequences mirrored along the vertical axis, for both training and testing (in practice, the extracted feature maps are mirrored to save computations). We always use *all* possible (overlapping) snippets of a certain length as data, both for training and testing (so for example a video of 27 frames yields 23 snippets of length $L = 5$). In each run, both the classifier *and* the set of templates are re-trained (although the template sets for different training sets are nearly identical except for occasional changes in the order of the basis vectors). The parameter settings given in the previous section were kept unchanged for all reported experiments.

5.1 Contributions of form and motion features

A main strength of the presented approach, compared to most other methods, is that it exploits dense form *and* motion features. A natural question therefore is, whether this is necessary, and how to combine the two. We have run experiments with our system, in which we have changed the relative weight λ of the two cues, or turned one of them off completely. The combination of form and motion consistently outperforms both form alone and motion alone, in all experiments we have conducted. Furthermore, the optimal relative weight turned out to be approximately the same across different data sets, and for snippets of different length (for our normalised similarity scores $\lambda = 0.7$, but being a scale normalisation between two independently computed feature sets, the value may depend on both the implementation and the parameters of feature extraction). For some datasets, form alone is a better cue than flow, while for others it is the other way round, depending on the set of actions and the recording conditions (see Fig. 5). This makes it unlikely that a strong bias towards one or the other pathway is introduced by our implementation, and supports the claim that explicitly extracting both cues increases performance.

5.2 How many frames?

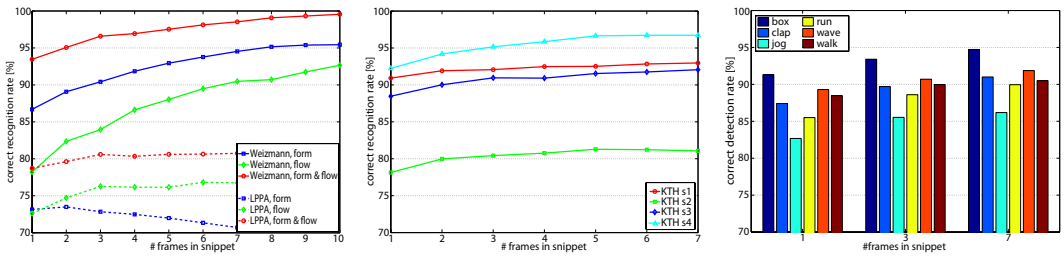

Figure 7: *Performance for different snippet lengths.* Left: WEIZMANN *and* LPPA *databases.* Middle: KTH *by scenario. Each scenario is trained and tested separately.* Right: *Per-category classification rates for increasing snippet length L (average over all* KTH *scenarios).*

Categorisation results of our method are shown in Fig. 6 and Fig. 7. For WEIZMANN, even $L = 1$ achieves 93.5% correct classification, snippets of ≥ 3 frames yield essentially perfect categorisation (< 1 wrong snippet per sequence). For LPPA, the result also saturates at $L = 3$. Note that for this dataset, the performance using only form *decreases* with growing snippet length—as more frames are added, the less schematic categories produce a larger variety of patterns, which becomes harder to capture with a small number of templates. In the KTH data, outdoor scenarios are more difficult than indoor (S4), because of extreme lighting variations. S2 performs worst, because we have no dedicated mechanism for scale invariance. Snippets of > 5 frames bring very little improvement. For the LPPA database. note that the "gold standard" achieved by human observers in the psychophysics experiment was 85% correct classification, meaning that the system's performance is 92.5% of the gold standard at $L = 1$, and 95% at $L = 7$. Longer snippets slightly increase performance, but the required length is not class-specific: the same classes are "easy", respectively "difficult",

independent of the snippet length.

	correct	frames		correct	frames
BLANK	99.6 %	10 / 10	*snippet1*	93.5 %	1 / 1
NIEBLES	55.0 %	1 / 12	*snippet3*	96.6 %	3 / 3
JHUANG	93.8 %	1 / 9	*snippet7*	98.5 %	7 / 7
			snippet10	99.6 %	10 / 10

Table 2: *Comparison with other results at snippet level (*WEIZMANN *database). The last column indicates the number of frames in a snippet (the unit for classification), and the number of frames used to compute features.*

Table 2 gives a comparison to other methods operating on single frames or snippets. Note that there are two groups using different paradigms, which cannot be directly compared. Our method, as well as Blank et al. (2005), look at snippets as atomic units, and assigns a label to a snippet. The methods of Jhuang et al. (2007); Niebles and Fei-Fei (2007) use a temporal window to compute features, but label only the central frame of the window. So for example, BLANK assigns a label to each snippet of 10 frames, using features computed on those 10 frames, whereas JHUANG assigns a label to every frame, using features computed in a 9-frame window. The first approach has the advantage that it does not require temporal look-ahead. Note especially the high classification rates even at snippet length $L = 1$ frame. The confusions, which do occur, are mainly between visually similar classes, such as *jogging/walking*, or *hand-clapping/hand-waving*. See confusion matrices in Fig. 6.

	snippet 1	*snippet 7*	entire seq.
KTH *all-in-one*	88.0 %	**90.9** %	81.5 % [NIEBLES]
KTH S1	90.9 %	93.0 %	**96.0** % [JHUANG]
KTH S2	78.1 %	81.1 %	**86.1** % [JHUANG]
KTH S3	88.5 %	**92.1** %	89.8 % [JHUANG]
KTH S4	92.2 %	**96.7** %	94.8 % [JHUANG]
WEIZMANN	93.5 %	98.6 %	**100.0** % [BLANK]
LPPA	78.7 %	**80.7** %	—

Table 3: *Categorisation results using snippets of different lengths, and comparison with published results for* whole sequences*. For* KTH S2*, note that we have no mechanism for scale invariance.*

Furthermore, we compare categorisation with snippets to the available results at *sequence* level, see Table 3. At $L = 7$ frames (less than 0.3 seconds of video), our results are comparable to the best ones obtained with full video sequences—in several cases, they are even better. The comparison confirms the message that a handful of frames are almost as informative for categorisation as the entire video.

5.3 Good and bad snippets

Clearly, not all parts of a motion sequence are equally suited for recognition. In a categorical (i.e., discriminative) framework this translates to the fact that certain snippets are easily miss-classified, because similar snippets occur in more than one category. Consequently, which snippets are distinctive for a certain category depends not only on the category itself, but on the entire set of "competing" categories as well. Note that the discriminative approach does not create this dependency, but only makes it apparent: with a generative model, one can measure the ability to reconstruct a pattern, independent of other categories. However, this does not change the fact that the model could reconstruct a very similar pattern for a different category.

When using a labelling algorithm made up of binary classifiers (such as our *one-vs-all* SVM), one has direct access to the "good" and "bad" snippets of each category: the most useful ones are those, which have a large margin (distance from the decision boundary), while the most ambiguous ones are those, which come to lie close to, or on the wrong side of, the decision boundary. By analysing the classifier output, one can therefore determine the usefulness of different snippets for the task.

Fig. 8 shows a few examples for the WEIZMANN dataset, computed at snippet length $L = 1$. For each class, we show the output of the binary *class-vs-nonclass* classifier on all frames (in-class samples in green, out-of-class samples in red). Furthermore we show: the least distinctive in-class frames for four different subjects (purple), the most confusing out-of-class frames for four different subjects (cyan), and the frames with the largest positive and negative margins, corresponding to the most confident decisions.[2] One can clearly see that the confusions are indeed caused by ambiguous body poses, such as similar parts of the *walking* and *running* cycle, and similar gestures, such as the arm movement of *jumping jack* and *waving both hands*.

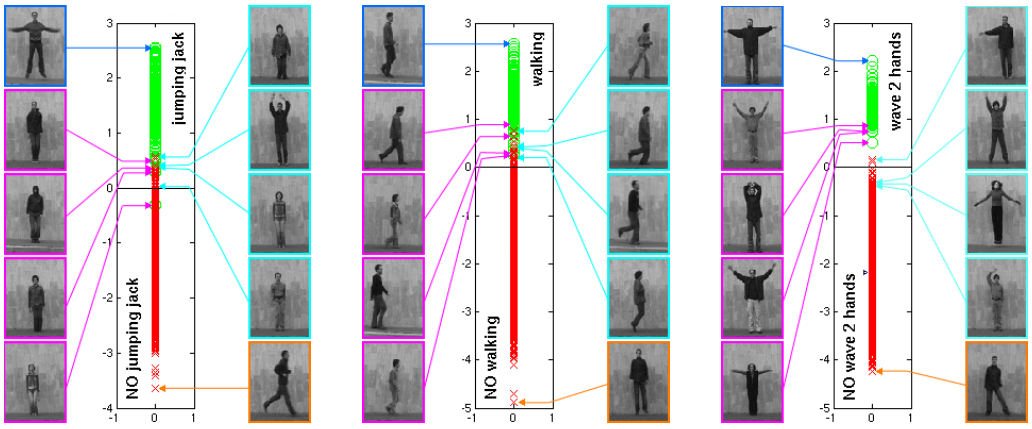

Figure 8: *Discriminative power of different frames for example classes* jumping jack*,* walking*, and* waving both hands*. See text for details.*

[2]Gaps between the displayed data points are from nearby, similar frames of the same subjects.

5.4 Comparison at sequence level

Most results in the literature are reported at the level of correctly classified sequences. To put our method in context, we therefore also compare it to the state-of-the-art at sequence level. Like other frame-based methods, we simply run our algorithm for all frames (snippets of length $L = 1$) of a sequence, and convert their individual labels to a sequence label through majority voting (a simplistic "bag-of-frames" model). The results are given in Table 4. Although our method was not specifically designed for this application, it achieves top performance on both data sets. This demonstrates the power of using dense form and motion.

KTH *all-in-one*		WEIZMANN	
bag-of-*snippet1*	**92.7** %	bag-of-*snippet1*	**100.0** %
NIEBLES	81.5 %	BLANK	**100.0** %
DOLLÁR	81.2 %	JHUANG	98.8 %
SCHÜLDT	71.7 %	WANG	97.8 %
JHUANG	91.7 %	ALI	92.6 %
(average of scenarios s1–s4, trained and tested separately)		DOLLÁR	86.7 %
		NIEBLES	72.8 %

Table 4: *Comparison of classification results at sequence level.*

The comparison should be taken with a grain of salt: in the action recognition literature, there is no established testing protocol, and different researchers have used varying sizes of training and test sets, different ways of averaging over runs, e.t.c. We always quote the best results someone has achieved. Still, the comparison remains indicative.

6 CONCLUSION

We have presented a method for human motion categorisation, which uses both form and motion features sampled densely over the image plane. The method was employed to experimentally investigate the question, how long video snippets need to be in order to serve as basic units for categorisation. In a detailed experimental evaluation, we have confirmed the advantage of explicitly extracting both form and motion cues. Furthermore, it has been shown that the method performs well on different databases without parameter changes, and that it matches the state-of-the-art, using fewer frames and no look-ahead. A main message of the study is that basic motion can be recognised well even with very short snippets of 1-7 frames (at frame rate 25 Hertz), as anticipated from the observation of biological vision.

A limitation of our current system is that it does not incorporate mechanisms for invariance to scale, rotation, and viewpoint (although it successfully handles scale changes up to a factor of $\approx$2, and viewpoint changes up to $\approx 30^\circ$, which are present in the KTH database). An open research question, which needs to be addressed before motion categorisation can be applied to realistic problems, is what the right "basic units" of human motion are, and how complex motion patterns—and ultimately unscripted human behaviour—can be represented as sequential or hierarchical combinations of such basic units. Another open issue is how to extend categorisation to scenarios with large numbers of categories, where it is no

longer feasible to independently learn a classifier for each category.

Acknowledgements

We would like to thank Hueihan Jhuang for providing unpublished per-frame results of her experiments, and for the accompanying explanation and discussion. Lydia Yahia-Cherif (LPPA, Collège de France) has kindly provided the LPPA dataset. This work has been supported by EU FP6 projects COBOL (NEST-043403) and DIRAC (IST-027787).

References

Ali, S., Basharat, A., and Shah, M. (2007). Chaotic invariants for human action recognition. In *Proc. 11th International Conference on Computer Vision, Rio de Janeiro, Brasil*.

Beintema, J. A. and Lappe, M. (2002). Perception of biological motion without local image motion. *Proc. National Academy of Sciences of the USA*, 99:5661–5663.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as space-time shapes. In *Proc. 10th International Conference on Computer Vision, Beijing, China*.

Carlsson, S. and Sullivan, J. (2001). Action recognition by shape matching to key frames. In *Proc. Workshop on Models versus Exemplars in Computer Vision*.

Casile, A. and Giese, M. A. (2005). Critical features for the recognition of biological motion. *Journal of Vision*, 5:348–360.

Comaniciu, D., Ramesh, V., and Meer, P. (2003). Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–575.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proc. 10th International Conference on Computer Vision, Beijing, China*, pages 886–893.

Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *Workshop on Performance Evaluation of Tracking and Surveillance (VS-PETS)*.

Efros, A. A., Berg, A. C., Mori, G., and Malik, J. (2003). Recognizing action at a distance. In *Proc. 9th International Conference on Computer Vision, Nice, France*.

Fanti, C., Zelnik-Manor, L., and Perona, P. (2005). Hybrid models for human motion recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA*.

Felleman, D. J. and van Essen, D. C. (1991). Distributed hierarchical processing in the primate visual cortex. *Cerebral Cortex*, 1:1–47.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of The Optical Society of America A*, 4(12):2379–2394.

Forsyth, D. A. and Ponce, J. (2003). *Computer Vision—A Modern Approach*. Prentice Hall Inc.

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for mechanisms of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202.

Gawne, T. J. and Martin, J. (2002). Response of primate visual cortical V4 neurons to two simultaneously presented stimuli. *Journal of Neurophysiology*, 88:1128–1135.

Giese, M. A. and Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Neuroscience*, 4:179–192.

Harnad, S., editor (1987). *Categorical perception: the groundwork of cognition*. Cambridge University Press.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160:106–154.

Jhuang, H., Serre, T., Wolf, L., and Poggio, T. (2007). A biologically inspired system for action recognition. In *Proc. 11th International Conference on Computer Vision, Rio de Janeiro, Brasil*.

Lampl, I., Ferster, D., Poggio, T., and Riesenhuber, M. (2004). Intracellular measurements of spatial integration and the max operation in complex cells of the cat primary visual cortex. *Journal of Neurophysiology*, 92:2704–2713.

Laptev, I. and Lindeberg, T. (2003). Local descriptors for spatio-temporal recognition. In *Proc. 9th International Conference on Computer Vision, Nice, France*.

Niebles, J. C. and Fei-Fei, L. (2007). A hierarchical model of shape and appearance for human action classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN*.

Niebles, J. C., Wang, H., and Fei-Fei, L. (2006). Unsupervised learning of human action categories using spatio-temporal words. In *Proc. 17th British Machine Vision Conference, Edinburgh, UK*.

Pontil, M. and Verri, A. (1998). Support vector machines for 3d object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(6):637–646.

Rao, C., Yilmaz, A., and Shah, M. (2002). View-invariant representation and recognition of actions. *International Journal of Computer Vision*, 50(2):203–226.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025.

Schüldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: a local SVM approach. In *Proc. International Conference on Pattern Recognition, Cambridge, UK*.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–426.

Serre, T., Wolf, L., and Poggio, T. (2005). Object recognition with features inspired by visual cortex. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA*.

Shawe-Taylor, J. and Christianini, N. (2000). *Support Vector Machines*. Cambridge University Press.

Wang, L. and Suter, D. (2007). Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical model. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN*.

Yacoob, Y. and Black, M. J. (1999). Parameterized modeling and recognition of activities. *Computer Vision and Image Understanding*, 72(2):232–247.