Action Snippets: How many frames does human action recognition require?

Konrad Schindler †
† BIWI, ETH Zürich, Switzerland
konrads@vision.ee.ethz.ch

Luc van Gool †,§
§ ESAT, KU Leuven, Belgium
vangool@vision.ee.ethz.ch

Abstract

Visual recognition of human actions 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 action label, or use relatively large look-ahead to classify each frame. Contrary to these strategies, human vision proves that simple actions can be recognised almost instantaneously. In this paper, we present a system for action recognition from very short sequences ("snippets") of 1–10 frames, and systematically evaluate it on standard data sets. It turns out that even local shape and optic flow for a single frame are enough to achieve ≈ 90% correct recognitions, and snippets of 5-7 frames (0.3-0.5 seconds of video) are enough to achieve a performance similar to the one obtainable with the entire video sequence.

1. Introduction

Recognising human actions in monocular video is an important scene understanding capability for applications such as surveillance, content-based video search, and human-computer interaction.

Past research in this domain can be roughly classified into two approaches: one that extracts a *global* feature set from a video [1, 9, 18, 27], and aims to assign a single label to the *entire video*, using these features. This paradigm obviously requires that the observed action 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* action label to each frame [3, 10, 17, 19]. If required, a global label for the sequence is usually obtained by simple voting mechanisms. The features are obtained by analysing a temporal window centred at the current frame, therefore the classification lags behind the observation, because a frame can only be classified after all frames in the temporal window have been observed.

Both approaches have achieved remarkable results, but human recognition performance suggests that they might bee using more information than required: we can correctly recognise actions from very short sequences (often even from single frames), and without temporal look-ahead.

1.1. Contributions

The question we seek to answer in this paper is ***how many frames are required to perform action recognition?*** 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 3). However, its answer has wide-ranging implications. Therefore, our goal is to establish a baseline, how long we need to observe a basic action, such as *walking* or *jumping*, in order to recognise it, if we try to use all available information.

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 displayed in Figure 1. The main message of our study is that ***very short snippets (1-7 frames), are sufficient for basic action recognition, with rapidly diminishing returns, as more frames are added***.

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's behaviour changes 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.

Figure 1. Examples of actions from databases WEIZMANN (top) and KTH (bottom). Note that even a single frame is often sufficient to recognise what a person is doing.

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

Furthermore, short snippets enable shorter processing times and rapid attention switching, in order to deal with further subjects or additional visual tasks, before they become obsolete.

We present a causal action recognition 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 the form and motion channels to learnt templates. The similarity scores are concatenated to a single feature vector and passed on to a classifier.

As far as we know, this is also the first practical implementation of a "biologically inspired" system, which is complete in the sense that it has a form as well as a motion pathway. The strategy to process form and motion independently, but with a similar sequence of operations, was inspired by the seminal work of Giese and Poggio [16]. 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 on two standard data sets, we evaluate the effect of changing the snippet length, and the influence of form and motion features. We also compare to other methods, both at the level of snippets and whole sequences, and obtain results on par with or better than the state of the art.

1.2. Defining "Action"

In the previous section, we have used the term *basic* action, to describe the units, into which human behaviour 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 an action. 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 of the action (*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 action decomposition: can, and should, complex actions be decomposed into sequences of simpler "atomic actions", which again can be recognised quickly?

The decomposition problem, which appears to be application-dependent, is *not* the topic of this study. We assume that a relatively small set of basic actions, such as *walking* or *waving*, form the set of possible action labels, and that the labels are relatively unambiguous (the most subtle difference we take into account is between *running* and *jogging*). 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 paper).

1.3. Action Snippets

The aim of the present work is not only to introduce yet another action classification method, but also to systematically investigate, how much information needs to be accumulated over time to enable action 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 seven images and seven flow fields (not six).

As will be demonstrated in section 4, using both shape and flow yields a marked improvement in recognition performance, compared to shape alone, or flow alone.

2. Related Work

Early attempts at human action recognition used the tracks of a person's body parts as input features [11, 22, 28]. This representation is an obvious choice, because physically the articulated motion is what defines an action. 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 cast action recognition as a shape matching problem [4]. An action is represented by a single unique pose, and recognition 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*. [10] 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*. [17] have extended the static scene recog-

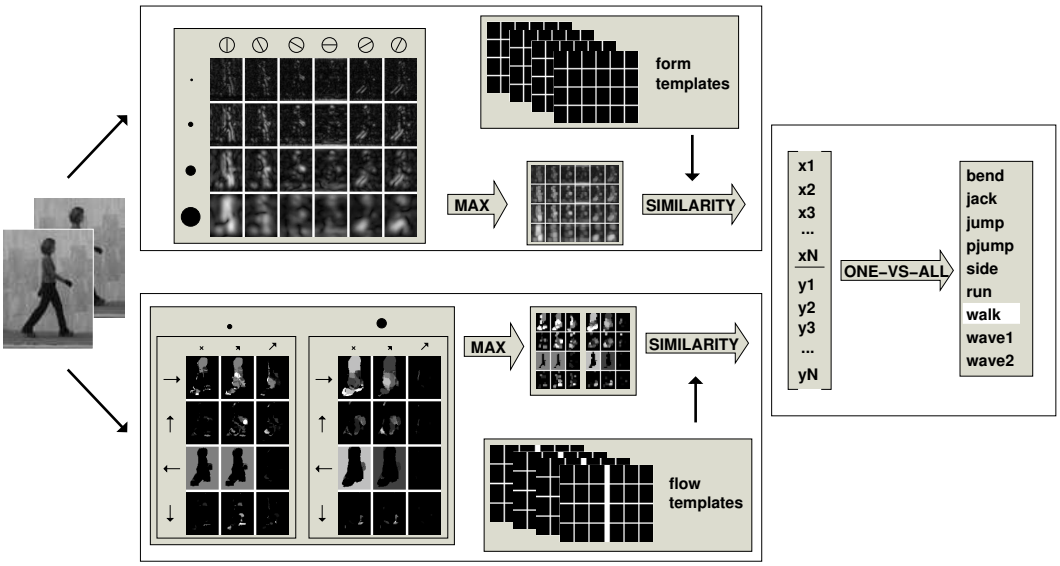

Figure 2. Overview of the recognition system. From a snippet, features are extracted in two parallel processing streams. In the form pathway (top), Gabor filters at multiple orientations and scales are applied. The motion pathway (bottom) extracts optic flow at different scales, directions, and speeds. In both pathways, the filter responses are MAX-pooled, and compared to a set of learnt templates. The similarities from both pathways are concatenated to a feature vector, and classified with a bank of linear classifiers.

nition model [25], by replacing form features with motion features. Like [10], 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 Li [19] 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 [9]. A constellation model for the features is learnt, and used to train a discriminative classifier.

Laptev and Lindeberg [18] 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 [18], or with a SVM [24].

Dollar *et al*. [9] present a different spatio-temporal interest point detector based on 1D Gabor filters, essentially searching for regions with sudden, or periodic, intensity changes in 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 [20].

Blank *et al*. [3] 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 also use silhouettes to classify at the sequence level [27]. 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*. [1] return to an articulated model, but follow only the main joints to make tracking more robust. Skeletonization is applied to silhouettes to obtain 2D stick figures, and their main joints are connected to joint trajectories. A sequence is represented by a set of chaotic invariants of these trajectories, and classified based on exemplars with a kNN-classifier.

3. Recognition Method

In order to make the best use of the available information, we explicitly extract both the object shape in each frame, and the optic flow between frames. Dense form (shape) and motion (flow) features are processed in what is sometimes called a "biologically inspired" manner, due to the similarity with the ventral and dorsal pathways of the visual cortex [12, 16]: the two types of information are pro-

cessed separately to yield two sets of high-level features, which are then merged before the final classification stage. Figure 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.* [5], but some researchers are of the opinion, that only form information from a number of key-frames is required, *e.g.* [2]. The key-frame paradigm has also been explored in machine vision [4]. Our experiments support the first view: using form and motion features consistently improves recognition performance, at least with our system architecture.

3.1. Input Data

Similar to [10, 17], we use a primitive attention mechanism: our input is a sequence of fixed-size image windows, centred at the person of interest. No foreground segmentation (silhouette) is required. Note that there is a subtle difference to [10]: they are interested in windows, for which the background can be considered uniform, so that only the relative articulations of the body are extracted. Although we (and also [17]) use a person-centred coordinate frame, too, our method does "see" a persons motion through the global image coordinate frame, by observing the inverse flow of the background within the stabilised window.

Compared to silhouette-based approaches, bounding boxes are more general. In particular, reliable silhouette extraction in practice requires a static background. On the contrary, bounding boxes are naturally obtained from person detectors based on sliding windows, *e.g.* [7], and/or trackers based on rectangular windows, *e.g.* [6].

3.2. Feature Extraction

Form features. Local shape is extracted from each frame separately. As descriptor, we use the responses of orientation filters, computed densely at every pixel. Specifically, we use a bank of log-Gabor filters, which allow a better coverage of the spectrum than standard (linear) Gabor filters with fewer scales [13]. The response g at position (x,y) and spatial frequency w is

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

with μ the preferred frequency of the filter, and σ a constant, which is set to achieve even coverage of the spectrum. The filter gain is proportional to the frequency, to compensate for the frequency spectrum of natural images, and give all scales equal importance. Our filter bank has 6 equally spaced orientations $\phi \in \{0°, 30°, 60°, 90°, 120°, 150°\}$, and 4 scales $\mu \in \{2, 4, 8, 16\}$ pixels. Only the magnitude $\|g^w(x,y)\|$ is used as shape descriptor, while the phase is discarded.

To increase robustness to translations and obtain a more compact representation, each orientation map is downsampled by MAX-pooling (sometimes called "winner-takes-all"). This operation was originally introduced in [14, 23], and has been shown to yield better translation invariance and feature preservation than linear averaging [14, 25]. 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 neighbourhood (receptive field) of the pixel (x,y). We sample the original maps at every 5$^{\text{th}}$ pixel, with a window of size (9×9). This size – determined experimentally – agrees with the findings of [17, 26]. Other than these works, we currently do not pool over scales.

In a last step, the orientation patterns are compared to a set of complex form templates learnt from examples (similar to [15, 17]), to yield a vector $\mathbf{q}_f$ of similarity scores. To learn an informative set of templates, the pooled orientation maps from training snippets are rearranged into one vector $\mathbf{h}$ per snippet, and simple linear PCA is applied. A fixed number of basis vectors $\{\mathbf{b}_i, i = 1 \ldots N\}$ are directly viewed as templates for relevant visual features.

To compute the similarity with the templates, the incoming vector $\mathbf{h}$ from a new image is scaled to norm 1 (a simple form of normalising the signal "energy"). In this way its projection $q_{f,i} = \langle \mathbf{h}, \mathbf{b}_i \rangle = \cos(\angle_{\mathbf{h}}^{\mathbf{b}_i})$ onto template $\mathbf{b}_i$ can be directly interpreted as a measure of similarity, where 1 means that the two are perfectly equal, and 0 means that they are maximally dissimilar. In our implementation, we use 500 templates (covering $\approx$40-50 % of the spectral energy, depending on the data set).

Note that when a snippet is more than one frame long, the features from all its frames are re-arranged into one single vector $\mathbf{h}$, implicitly using the frame ordering. A snippet of length L is therefore more informative than an unordered bag of L frames.

Motion features. Dense optic flow at every frame is computed by direct template matching to the previous frame, using the L_1-norm (sum of absolute differences). Although optic flow is notoriously noisy, we prefer not to apply any smoothing, for the following reasons: spatial smoothing blurs the flow field at discontinuities, where the motion information is important, especially if no figure-ground segmentation is available; smoothing over time is incompatible with the concept of independent snippets without temporal look-ahead.

To obtain a representation consistent with the log-Gabor maps for form, the optic flow is converted to a set of response maps for different "flow filters", each with different preferred flow direction and speed. A filter's response is maximal, if the direction and speed at a certain location exactly match the preferred values, and decreases linearly, as the direction and/or speed changes. The filter responses are computed at 2 spatial scales (window size $\psi \in \{8, 16\}$ pixels), 4 equally spaced directions (half-wave rectified,

$\phi \in \{0°, 90°, 180°, 270°\}$), and 3 scale-dependent velocities $(0, \psi/4, \psi/2)$.

The rest of the motion pathway works similarly to the form pathway, *i.e*. local flow maps are MAX-pooled, and all maps from a snippet are converted to a vector $\mathbf{q}_m$ of similarity values, by comparing to a set of complex flow templates learnt with PCA. Note that although we use unsmoothed optic flow as low-level feature, the templates are smooth due to the built-in denoising capabilities of PCA. The same parameters are used in the form and flow pathways (down-sampling factor $F = 5$, pooling size $W = 9$, number of templates $T = 500$).

3.3. Classifier

The feature vectors for form and motion are merged by simple concatenation, $\mathbf{q} = [(1-\lambda)\mathbf{q}_f \ , \ \lambda\mathbf{q}_m]$, resulting in a vector of dimension 1000. The optimal weighting factor $\lambda \in [0..1]$ has been determined experimentally, and has proved to be stable across different data sets and snippet lengths, see section 4.

As classifier for K action classes, we train a bank of K linear one-vs-all SVMs, each with weights $W = (K-1)$ for positive samples, and $W = 1$ for negative ones, to account for the uneven sample numbers. We purposely keep the classification part simple. One alternative would be to use a non-linear kernel for the SVM. However, this yield very little, if any, improvement (probably due to the high dimension of the data), while introducing additional parameters. Alternative multi-class extensions are the all-pairs strategy, or bitwise learning of error-correcting output codes [8]. In practice, all three strategies give similar results, with the one-vs-all method needing the smallest number of classifiers. Note also that it has the most obvious interpretation in terms of biological vision systems, with each binary classifier corresponding to a neural unit, which is activated by actions of a certain class.

3.4. Relation to existing methods

Since recent methods for action recognition, 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 [10, 17]. These three methods are less demanding than [1, 3, 27], which require a segmented silhouette, but more demanding than interest-point based methods [18, 19], 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 fall onto the person. No method has yet been tested in cluttered environments with many distractors.

In terms of features used, [10, 17] extract only optic flow, [3, 27] only silhouette shape, and [18] have used either form or motion, while our work, as well as [19], extract both cues independently.

More generally, our method belongs to a school, which favours densely sampled features over sparse interest points for recognition problems, *e.g*. [7, 21]. It can also be considered a biologically inspired model, with parallels to the "standard model" of the visual cortex [23]: 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, which each compare 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 action class.

The other model, which intentionally follows a biologically inspired architecture [17], 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 image), and are learnt by random sampling, while we apply PCA to find a template set, which is in some sense optimal, albeit less biologically plausible.

4. Experimental Evaluation

We use two data sets for our evaluation, which have become the de-facto standards for 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.* 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 set was recorded for [18, 24], and consists of 25 subjects performing 6 different actions: *boxing, hand-clapping, jogging, 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

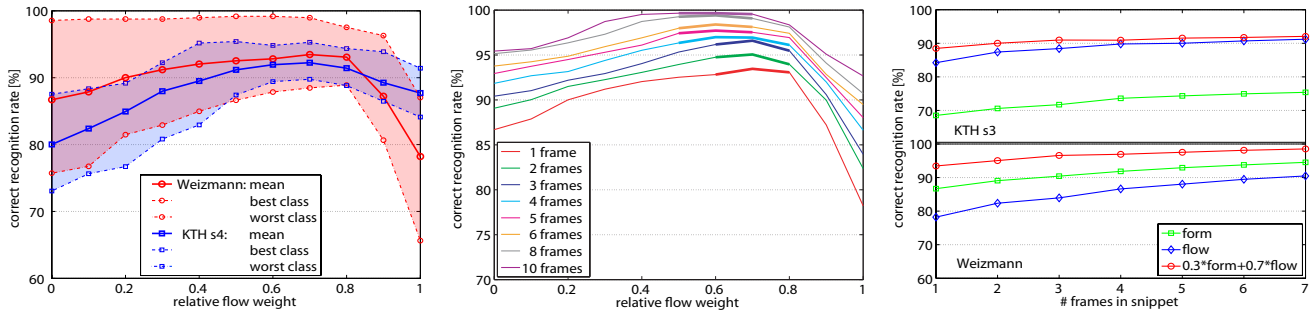

Figure 3. Influence of form and motion features on recognition performance. *Left:* Recognition rates with different relative weights of both pathways, computed on WEIZMANN and KTH S4 data bases. Shown are the correct recognition rates at snippet length $L = 1$, for the best and worst action classes, and the average over all classes. *Middle:* Recognition rates with different relative weights and different snippet lengths, shown for the WEIZMANN database. Bold lines show where the performance peaks. *Right:* Recognition rates using only form, only motion, or the best combination, shown for WEIZMANN and KTH S3 databases.

WEIZMANN

	bend	jack	jump	pjump	side	run	walk	wave1	wave2
bend	99.0	0.4		0.4	0.4			0.4	
jack		91.2		2.1					0.8
jump			93.0		4.7	2.3	3.3		
pjump		6.2	0.2	95.9		4.3		3.7	2.9
side		0.4	3.1	1.2	88.5	2.1	3.1		
run	0.2	0.4	2.7	0.4	0.4	89.9	2.1		
walk	0.6	0.4	1.0		6.0	1.2	91.6		
wave1	0.2	0.8				0.2		95.9	
wave2		0.2							96.3

KTH

	box	clap	jog	run	wave	walk
box	94.2	5.2	0.3	0.1	3.1	1.7
clap	2.8	89.0		0.2	6.5	0.2
jog	0.5		81.7	14.2	0.4	7.4
run	0.1		10.6	84.2		1.0
wave	1.2	5.0	0.1	0.1	89.1	0.1
walk	1.2	0.8	7.4	1.3	0.9	89.7

	correct	# frames
BLANK [3]	99.6 %	10 / 10
NIEBLES [19]	55.0 %	1 / 12
JHUANG [17]	93.8 %	1 / 9
SNIPPET 1	93.5 %	1 / 1
SNIPPET 3	96.6 %	3 / 3
SNIPPET 7	98.5 %	7 / 7
SNIPPET 10	99.6 %	10 / 10

Table 1. Recognition results at snippet level *Left, Middle:* Confusion matrices for $L = 1$ (form and motion at a single frame); the main confusions are walking/jogging, and running/jogging. *Right:* Comparison with other methods (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.

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.

To account for the symmetry of human actions, 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 $L = 5$). The parameter settings given in the previous section were kept unchanged for all reported experiments.

4.1. Contributions of form and motion features

A main strength of the presented approach, compared to other action recognition 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 system $\lambda = 0.7$, but being a scale normalisation between two independently computed feature sets, the value depends on both the implementation and the parameters of feature extraction).

It is also interesting to note that for some data sets, form alone is a better cue than flow alone, while for others it is the other way round. This makes it unlikely that a strong bias towards one or the other pathway is introduced by our specific implementation, and supports the claim that explicitly extracting both cues increases recognition performance. See Figure 3 for details.

4.2. How many frames?

Recognition results of our method are shown in Figure 4, and a comparison to other recognition methods operating on single frames or snippets is given in Table 1. Note that

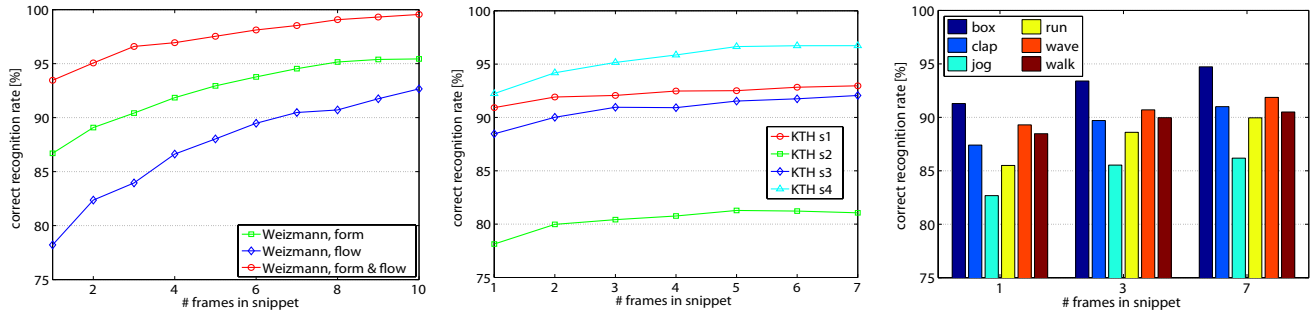

Figure 4. Performance for different snippet lengths. *Left:* WEIZMANN database. Even $L=1$ achieves 93.5% correct recognitions, snippets of ≥ 3 frames yield essentially perfect recognition (<1 wrong snippet per sequence). *Middle:* KTH by scenario. Each scenario is trained and tested separately. Outdoor 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. *Right:* Per-class recognition rates for increasing snippet length L (average over all KTH scenarios). 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.

there are two groups using different paradigms, which cannot be directly compared. Our method, as well as [3], looks at snippets as atomic units, and assigns a label to a snippet. The methods [17, 19] 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 recognition rates even at snippet length $L = 1$ frame. The confusions, which do occur, make sense – miss-classifications happen mainly between similar classes, such as *jogging–walking*, or *handclapping–handwaving*. See confusion matrices in Table 1.

Furthermore, we compare recognition with snippets to the available recognition results at *sequence* level, see Table 2. 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 short action snippets with a handful of frames are almost as informative as the entire video.

	SNIPPET 1	SNIPPET 7	entire seq.
KTH *all-in-one*	88.0 %	**90.9** %	81.5 % [20]
KTH S1	90.9 %	93.0 %	**96.0** % [17]
KTH S2	78.1 %	81.1 %	**86.1** % [17]
KTH S3	88.5 %	**92.1** %	89.8 % [17]
KTH S4	92.2 %	**96.7** %	94.8 % [17]
WEIZMANN	93.5 %	98.6 %	**100.0** % [3]

Table 2. Comparison of results using snippets with best published results using *whole sequences*. For KTH S2, note that other than our system, [17] has a mechanism for scale invariance.

4.3. 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 3.

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, *etc*. We always quote the best results someone has achieved. Still, the comparison remains indicative.

5. Conclusion

We have presented a method for human action recognition, which uses both form and motion features sampled densely over the image plane. The method was em-

KTH *all-in-one*		WEIZMANN	
bag-of-SNIP_1	**92.7** %	bag-of-SNIP_1	**100.0** %
JHUANG† [17]	91.7 %	BLANK [3]	**100.0** %
NIEBLES [20]	81.5 %	JHUANG [17]	98.8 %
DOLLÁR [9]	81.2 %	WANG [27]	97.8 %
SCHÜLDT [24]	71.7 %	ALI [1]	92.6 %
†*Average of scenarios s1–s4, trained and tested separately.*		DOLLÁR [17]	86.7 %
		NIEBLES [19]	72.8 %

Table 3. Comparison of recognition results at sequence level. Although our method was not designed for this application, it achieves top performance on both data sets, with a simple "bag-of-snippets". This result demonstrates the power of using both form and motion information.

ployed to experimentally investigate the question, how long video snippets need to be, to serve as basic units for action recognition. 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 any 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 actions 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 systems.

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 ≈ 2, and viewpoint changes up to $\approx 30^\circ$, which are present in the KTH database). Furthermore, it remains to be investigated, how classification errors are distributed within an action class, or in other words, which snippets within a motion sequence are particularly suitable (or unsuitable) as "key snippets" to recognise the action.

An open research question, which needs to be addressed before action recognition can be applied to realistic problems, is what the right "basic units" of human action are, and how complex actions – and ultimately unscripted human behaviour – can be represented as sequential or hierarchical combinations of such basic units.

Acknowledgements

We would like to thank Hueihan Jhuang for providing unpublished per-frame results of her experiments, and for the accompanying explanation and discussion. This work has been supported by EU projects COBOL (NEST-043403) and DIRAC (IST-027787).

References

[1] S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. In *Proc ICCV*, 2007.

[2] J. A. Beintema and M. Lappe. Perception of biological motion without local image motion. *P Natl Acad Sci USA*, 99:5661–5663, 2002.

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Proc ICCV*, 2005.

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

[5] A. Casile and M. A. Giese. Critical features for the recognition of biological motion. *J Vision*, 5:348–360, 2005.

[6] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE T Pattern Anal*, 25(5):564–575, 2003.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc ICCV*, pages 886–893, 2005.

[8] T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *J Artif Intell Res*, 2:263–286, 1995.

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

[10] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proc ICCV*, 2003.

[11] C. Fanti, L. Zelnik-Manor, and P. Perona. Hybrid models for human motion recognition. In *Proc CVPR*, 2005.

[12] D. J. Felleman and D. C. van Essen. Distributed hierarchical processing in the primate visual cortex. *Cereb Cortex*, 1:1–47, 1991.

[13] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J Opt Soc Am A*, 4(12):2379–2394, 1987.

[14] K. Fukushima. Neocognitron: a self-organizing neural network model for mechanisms of pattern recognition unaffected by shift in position. *Biol Cybern*, 36:193–202, 1980.

[15] M. A. Giese and D. A. Leopold. Physiologically inspired neural model for the encoding of face spaces. *Neurocomputing*, 65-66:93–101, 2005.

[16] M. A. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements. *Nat Neurosci*, 4:179–192, 2003.

[17] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *Proc ICCV*, 2007.

[18] I. Laptev and T. Lindeberg. Local descriptors for spatio-temporal recognition. In *Proc ICCV*, 2003.

[19] J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *Proc CVPR*, 2007.

[20] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatio-temporal words. In *Proc BMVC*, 2006.

[21] M. Pontil and A. Verri. Support vector machines for 3d object recognition. *IEEE T Pattern Anal*, 20(6):637–646, 1998.

[22] C. Rao, A. Yilmaz, and M. Shah. View-invariant representation and recognition of actions. *Int J Comput Vision*, 50(2):203–226, 2002.

[23] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nat Neurosci*, 2:1019–1025, 1999.

[24] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *Proc ICPR*, 2004.

[25] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Object recognition with cortex-like mechanisms. *IEEE T Pattern Anal*, 29(3):411–426, 2007.

[26] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *Proc CVPR*, 2005.

[27] L. Wang and D. Suter. Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical model. In *Proc CVPR*, 2007.

[28] Y. Yacoob and M. J. Black. Parameterized modeling and recognition of activities. *Comput Vis Image Und*, 72(2):232–247, 1999.