VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

HUMAN ACTION RECOGNITION IN VIDEO

DISERTAČNÍ PRÁCE
PH.D. THESIS

AUTOR PRÁCE
AUTHOR

Ing. IVO ŘEZNÍČEK

BRNO 2014

BRNO UNIVERSITY OF TECHNOLOGY

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER
GRAPHICS AND MULTIMEDIA

HUMAN ACTION RECOGNITION IN VIDEO

PH.D. THESIS

AUTHOR Ing. IVO ŘEZNÍČEK

SUPERVISOR prof. Dr. Ing. PAVEL ZEMČÍK

BRNO 2014

Abstract

This thesis focuses on the improvement of human action recognition systems. It reviews the state-of-the-art in the field of action recognition from video. It describes techniques of digital image and video capture, and explains computer representations of image and video. This thesis further describes how local feature vectors and local space-time feature vectors are used, and how captured data is prepared for further analysis, such as classification methods. This is typically done with video segments of arbitrarily varying length. The key contribution of this work explores the hypothesis that the analysis of different types of actions requires different segment lenghts to achieve optimal quality of recognition. An algorithm to find these optimal lengths is proposed, implemented, and tested. Using this algorithm, the hypothesis was experimentally proven. It was also shown that by finding the optimal length, the prediction and classification power of current algorithms is improved upon. Supporting experiments, results, and proposed exploitations of these findings are presented.

Keywords

Optimal analysis length of action, local space-time features, bag-of-words repesentation, visual vocabulary, SVM.

Contents

Chapter 1

Introduction

Nowadays the world is heading towards to a total monitoring of everything what is going on. That induces a creation of systems for detection, let us say, anomalies that are physically happening around us. One of the needs is, for example, the monitoring of some public areas for violent behavior detection. This can be achieved by a number of humans that are located at monitored places and are able to prevent other beings from performing such activities. This solution is very expensive in terms of a huge number of humans that are performing the monitoring, and all of them are potentially in danger on a place.

Contemporary techniques give an opportunity to adapt above presented procedure and deploy video cameras to the places where monitoring needs to be performed, and the humans are observing the situation remotely. This solution reduces the number of observers on place, and the safety of such humans is higher than before. Such monitoring procedure is often called as surveillance.

As time has gone by the cameras' quality is higher and higher, the output signal can be digitized, and the possibility of processing of the camera signal by computers comes forward. Such processing can be split into two main areas; it is (1) still image processing, detection and localization of humans, faces, animals, or general objects and (2)

video processing, such as behavior detection in front of the camera, car accident detection, smoke detection, fight detection, etc. All those detections may be very valuable for the surveillance systems. It is highly needed to remove the human factor from surveillance application and let the computer detect all dangerous or unwanted situations that may happen but that situation did not come up yet.

Nowadays such computer detection techniques are being used as a support of the surveillance system's operator and he/she finally decides whether the detection is valid and further reactions need to be performed.

This work focuses on the human action recognition from digitized video streams but current research in such field generally approaches the problem in a way, that some new algorithm is proposed and afterwards tested on the standardized sets of video sequences and the solution's quality is then measured. All sets of video sequences generally contain totally variable video streams. The main problems can be seen (i) as the variable length of samples, (ii) on the fact that samples contain an amount of other actions and somewhere at the end or inside of the sample the wanted action is performed, (iii) the presented behavior is fully or partially done outside of video frames. Majority of researches use the whole example as one entity and above presented facts are simply not taken into account.

The main question remains; it is whether some, let us say, optimal analysis length of action can be found and consequently whether online detection system, which processes only such restricted length of the video, can be built and whether that system produces comparable results to the currently well known solutions.

In this thesis the action recognition solutions were investigated and the technique which leads to the state-of-the-art performance were replicated. The whole video processing may be understood as a series of transformation which are described in subsequent chapters. In Chapter 2 the image and video representations in a computer are described and extended processing pipeline including description of its components is presented as well. Chapter 3 presents the definition

of the *optimal* analysis length and the algorithm for obtaining of that property of human action recognition processing. The conclusion of the work is presented in Chapter 4.

Chapter 2

State of the art

The computer representation of an image and an video is usually done by using a 2D or 3D raster of pixels (picture elements). In its digital form, the image and video functionos are represented by a limited size arrays of discrete values and the fact that this representation accurately represents the usually continuous original image and video functons is due to the well known sampling theorem.

The arguments of image and video functions are position co-ordinates and in case of video also time. The functions can be seen as grayscale or color based on a color model used for image representation. The frequently used one is the RGB model, thus the functions return a separated value for each of the Red, Green, and Blue channels.

2.1 Image and video content representation

The content representation is further extracted from given image or video function. The possible way how to make this possible is to use the local features [11, 10, 1, 12] or the spatio-temporal features [7, 8, 16, 3, 15] in case of videos. These features are usually associated to a certain locations of given image of video function and are described by using feature vectors. These features are subsequently processed

and are used as an input of the classification process.

The feature vectors obtained from the images of videos usually have constant size but extractor produces varying count of feature vectors. These vectors are used in the *Vocabulary search* box. This block produces a set of nearest clusters related to particular input feature vector. From these sets a bag-of-words representation is constructed in the *Hist* box. The Bag-of-words [14] representation is one fixed-sized feature representation of input entity (image or video). This representation is used as an input to the *classifier box* which decides the type (class) of the input entity.

The above mentioned procedure requires a visual vocabulary creation, only then the whole processing is possible. Visual vocabulary is created (as shown in the upper part of Figure 2.1) from entities (images or videos) of the training set of a dataset. Those are transformed to feature vectors using the same *feature extract* box with the same settings as the one which is described above. The output feature vectors define the feature space which needs to be quantized and modelled using a clustering algorithm. The *Clustering* box constructs a visual vocabulary (or by other words, the quantized representation of its input feature space) which is latter used for bag-of-words representation creation. For such purposes the k-means [17] algorithm is usually used.

2.2 Extended processing pipeline

The extended pipeline which depicts the whole data processing is shown in Figure 2.1. The extension consists of an usage of multiple feature extractors [15, 16] as well as visual vocabularies, several classifiers and finally it should lead to the creation of better output classifier. This approach has generally a potential to improve the quality of a video processing solution because every solved problem has its own particularity and this processing can select well performing feature extractors for such purpose.

All those facts cause an usage of more of bag-of-words [14] units.

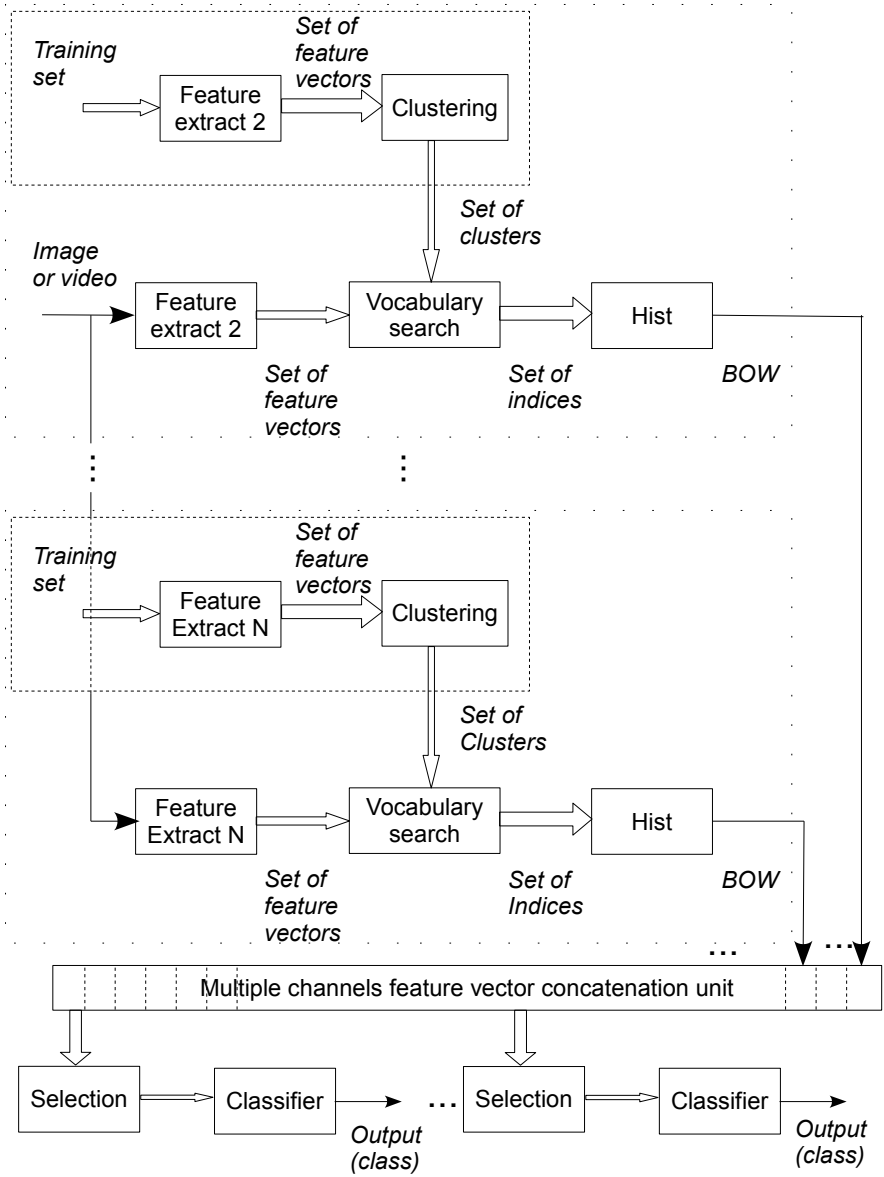

Figure 2.1: Extended pipeline schema

The outputs from bag-of-words units ale called a channels and are "concatenated" into one long multiple channels feature vector. Every element in this feature vector consists of one bag-of-words full representation which is not divided into smaller parts.

From the multiple channels feature vector a number of selections are constructed and from each selection a classifier is created. The term selection is considered as a "reduction" of number of channels, one extreme possibility is when all channels are selected, on the other side is when only one channel is selected. More classifiers are produced in this way and the best one needs to be selected.

As a classifier creation process the multi-kernel support vector machine with a multichannel gaussian kernel [6, 2] was used.

Selected best classifier uses only finite set of input channels which were selected in corresponding selection unit. Classifier produced by this pipeline should be used for prediction of unwanted entites, only corresponding feature extractors need to be provided for the computation (this can reduce the computation effort).

In this work, a validation dataset is used for best classifier selection.

Chapter 3

Optimal analysis length

Many scientific papers concerning human action recognition focus on the improving some part or some parts of the video processing pipeline. Currently, the improvements are mainly in the development of new local space-time features [7, 16, 3, 15] and in combining known space-time features [15, 5]. Other solutions where the processing pipeline is different also exist such as deep-learning techniques [9]. All proposed procedures have something in common; the use of a dataset for evaluating quality.

Datasets usually contain a number of videos with variable length, the proposed actions are usually located anywhere within a particular example. When using processing similar to the one presented in Chapter 2, the bag-of-words unit processes the whole video sequence which is potentially large and will have varying size. Such a procedure is called off-line processing.

Some scenarios require the usage of a processing step, in which the number of frames processed by the bag-of-words unit is fixed to a predefined value. In other words, the bag-of-words representation is extracted from a sub-shot of the video with constant length. This is called on-line processing.

The following sections are organized as follows, in Section 3.1 the main contribution of the thesis is presented. The main experiments

about on-line processing are presented in Section 3.2.

3.1 Determining the optimal length of the analyzed video

The contribution of this thesis is in: (i) to state the hypothesis that an *optimal* analysis length exists for on-line action recognition solutions and (ii) the proof of this hypothesis through an algorithmic solution.

The *optimal* length of analyzed video ℓ_o exists for each on-line human action recognition system where the solution has a quality q_o which is comparable to the off-line solution quality q, formally $q_o \geq q - \epsilon$ and this assumption applies for each arbitrarily small ϵ. For appropriate values of ϵ, the *optimal* analysis length of videos of certain actions can be much smaller than the potentially unrestricted length, which is processed by the off-line solution.

Subsequently, according to the hypothesis, for certain actions an analysis action length may exist, such that the on-line solution gives better quality than the off-line solution.

Such an algorithmic proof of the statements has two inputs:

(i) A dataset with start and end positions of each annotated action (DS),

(ii) A state-of-the-art method in which quality is expressed by a function $\mathbf{M}$: DS $\longrightarrow \mathbb{R}$.

Formally, the algorithm can be interpreted as a mapping:

$$(\mathbf{M}, DS, q, \epsilon) \longrightarrow ((\ell_o, q_o), (\ell_b, q_b)) \tag{3.1}$$

L represents all lengths to be analyzed (in number of frames)

$$L = \langle \ell_{min}, \ell_{max} \rangle \cap \mathbb{N} \tag{3.2}$$

Metrics of qualities for each length of sequence are represented by a vector D.

$$D = \{(\ell_i, q_i) \,|\, \ell_i \in L, q_i = \mathbf{M}(\mathbf{CO}(DS, \ell_i))\} \tag{3.3}$$

The *optimal* analysis length of the video is ℓ_o, and the analysis length where the maximum quality is achieved is ℓ_b. Solution quality q_o and q_b are defined as well.

$$(\ell_o, q_o) \in D \,|\forall (\ell_j, q_j) \in D, \ell_j < \ell_o, q_j < q - \epsilon) \tag{3.4}$$

$$(\ell_b, q_b) \in D \,|\forall (\ell_j, q_j) \in D, q_j \leq q_b \tag{3.5}$$

The function $\mathbf{CO}(DS, \ell_i)$ returns a dataset (which is based on DS) where each video sample has constant length ℓ_i.

It should be noted that it is not guaranteed that (l_o, q_o) exist but it does for most cases.

A practical way to obtain (l_o, q_o) and (l_b, q_b) is described in the algorithm below:

(1) Get ℓ_{min} and ℓ_{max}
(2) **for** i = ℓ_{min}:ℓ_{max} {
    D[i]= $\mathbf{M}(\mathbf{CO}(\mathrm{DS}, \mathrm{i}))$
    }
(3) perform:
    o = NULL; **for** i = ℓ_{min}:ℓ_{max} {
    if(D[o] $\geq$ (q - ϵ)) { o = i; **break**; }
    }
    and return ℓ_o = o and q_o = D[o].
(4) perform:
    b = ℓ_{min}; **for** i = ℓ_{min}:ℓ_{max} {
    if(D[i] > D[b]) { b = i; }
    }
    and return ℓ_b = b and q_b = D[b].

Algorithm 1: Verification algorithm pseudocode.

It should be noted that the algorithm is not optimal. The main reason why it has been constructed, is to prove the presented hypothesis, and to obtain the *optimal* analysis length. The creation of a more powerful algorithm is not the focus of this thesis.

The experimental proof of the hypothesis using this algorithm is shown below.

3.2 Optimal length experiments

The purpose of the experiments performed in this section is the verification of the hypothesis about the *optimal* analysis length proposed in Section 3.1.

We have used the pipeline presented in Chapter 2 and the Hollywood2 [8] dataset. This dataset contains twelve action classes from Hollywood movies, namely: *answering the phone, driving car, eating, fighting, getting out of the car, hand shaking, hugging, kissing, running, sitting down, sitting up and standing up.*

We investigated the recognition algorithm behavior in such a way that pieces of video containing the action were presented to the algorithm at randomly selected positions inside the actions. For example, the actions were known to start earlier than the beginning of the processed piece of video, and ended only after the end of the presented piece of video. For this purpose, we had to reannotate the Hollywood2 dataset (all three its parts - train, autotrain and test) to obtain precise beginning and ending frames of the actions.

In our experiments, we have been trying to depict a dependency between the length of video shot, being an input to the processing, and the accuracy of the output. We have set the minimum shot length to 5 frames, more precisely the 5 frames from which the space-time point features are extracted. The maximum shot length was set to 100 frames and the frame step was set to 5 frames.

The space-time features extractor process N previous and N consequent frames of the video sequence in order to evaluate the points of interest for a single frame. Therefore, 2*N should be added to every figure concerning the number of frames to get the total number of frames of the video sequence to be processed. In our case, N was equal to 4 so that, for example, the 5 frames processed in Figures 3.1, 3.2 and 3.3 mean 13 frames of the video.

A classifier has been constructed for every video shot length considered. The training samples were obtained from the training part of the dataset

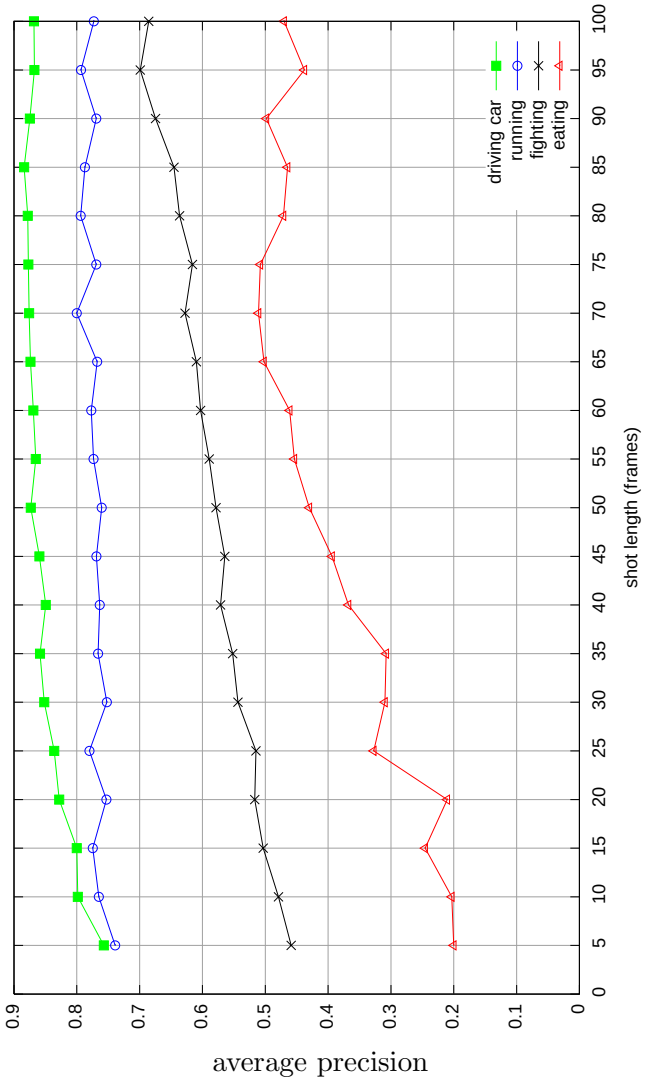

Figure 3.1: Dependency of the average precision on the length of the shot achieved for first group of classes contained in the Hollywood2 dataset.

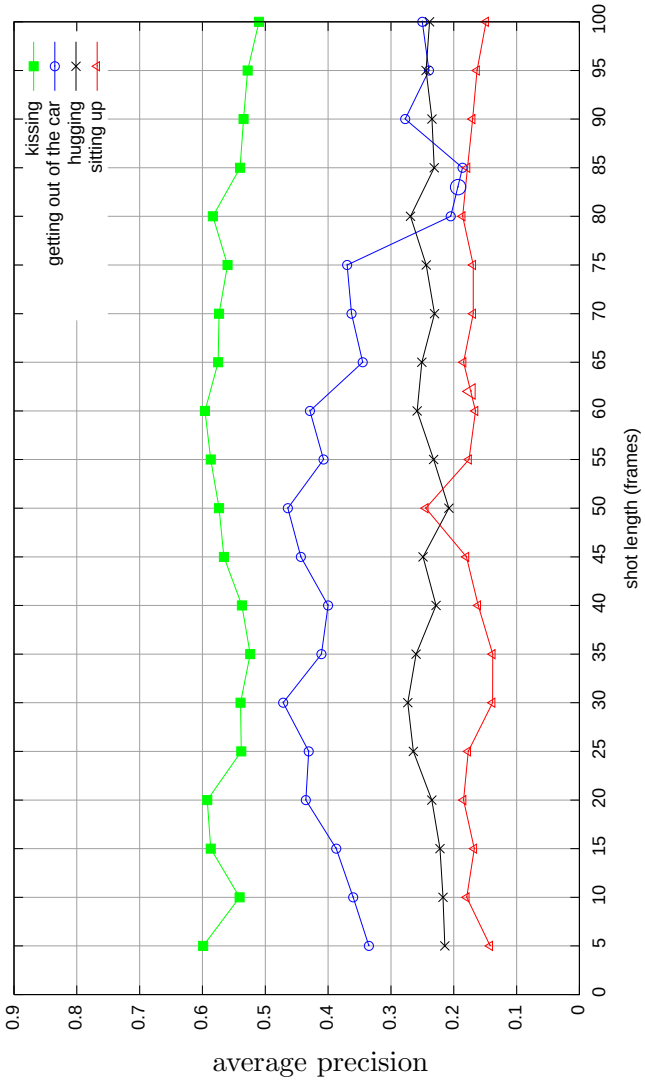

Figure 3.2: Dependency of the average precision on the length of the shot achieved for second group of classes contained in the Hollywood2 dataset. It should be noted that the big marks indicating the average action length shown in the charts for actions shorter than 100 frames.

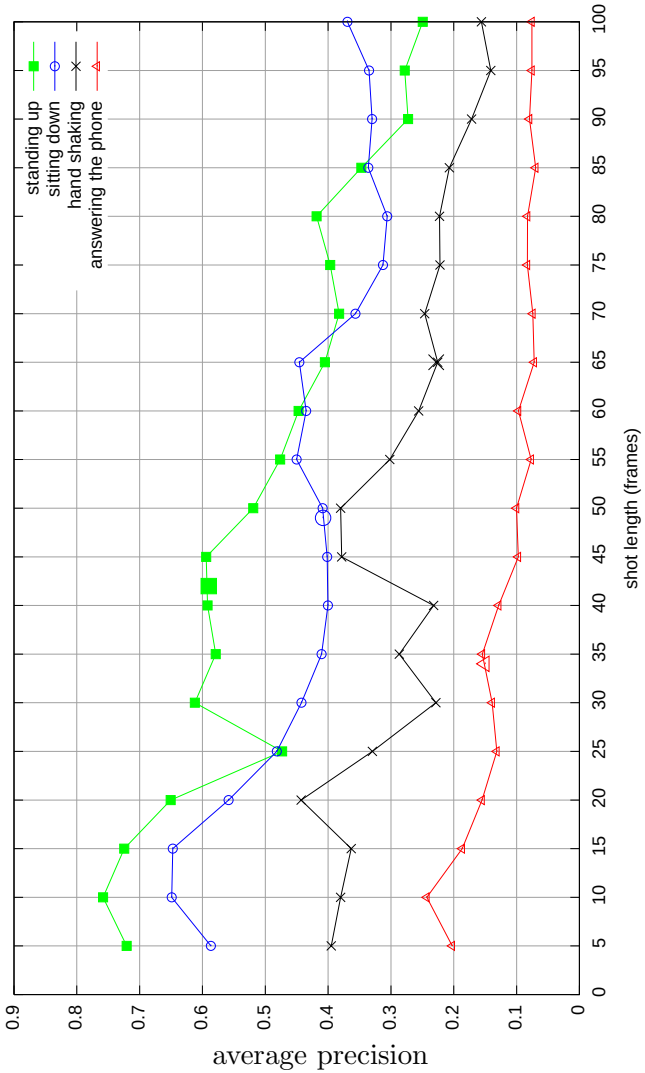

Figure 3.3: Dependency of the average precision on the length of the shot achieved for third group of classes contained in the Hollywood2 dataset. It should be noted that the big marks indicating the average action length shown in the charts for actions shorter than 100 frames.

Table 3.1: Results of the experiments. The first four columns show the accuracy (average precision) for the selected video sizes, the consequent column shows the reference accuracy reached for unrestricted video size, and the final column shows the minimum number of frames needed to achieve 90% of the precision achieved using the unrestricted video size. Column description: **A** – Unrestricted video size accuracy, **B** – Number of frames to achieve 90% accuracy

	Video size (frames)				**A**	**B**
Action	5	10	30	90		
driving car	0.757	0.798	0.852	0.874	0.848	10
running	0.739	0.765	0.752	0.769	0.812	10
fighting	0.459	0.479	0.543	0.675	0.718	90
eating	0.2	0.203	0.309	0.498	0.326	25
kissing	0.599	0.541	0.540	0.535	0.597	5
get out of the car	0.335	0.36	0.471	0.277	0.358	5
hugging	0.214	0.217	0.273	0.235	0.264	25
sitting up	0.142	0.179	0.138	0.17	0.163	10
standing up	*0.721*	*0.758*	*0.612*	*0.273*	*0.598*	*5*
sitting down	*0.587*	*0.649*	*0.442*	*0.33*	*0.654*	*10*
hand shaking	*0.395*	*0.38*	*0.229*	*0.172*	*0.232*	*5*
answer the phone	*0.201*	*0.243*	*0.139*	*0.08*	*0.225*	*10*

in the following way: the information of the start and stop position in the currently processed sample was used and large number of the randomly selected subshots were obtained. The training dataset has 823 video samples in total and from each sample, we extracted 6 subshots on average.

The actual evaluation of the classifier has been done four times in order to obtain the information about reliability of the solution. Also, the above mentioned publications used the 823 samples for evaluation purposes and we wanted our results to be directly comparable. The results shown in Table 3.1 and Figures 3.1, 3.2 and 3.3 present the average of the results of the four runs. For this purpose we have randomly determined a position of starting frame of a testing subshot within a testing sample four times. The above approach brings us two benefits - the final solution accuracy can be measured using an average precision metric and the results obtained through the testing can be easily comparable to the published state-of-the-art solutions. The results were compared with the accuracy achieved on the video sequences with completely unrestricted size that are close to the state-of-the-art [13].

The parameters for feature processing and classification purposes were as follows: the tested feature extractor is the dense trajectories extractor, which produces four types of descriptors, namely: HOG, HOF, DT and MBH. These four feature vectors were used separately. For each descriptor a vocabulary of 4000 words was produced using the k-means method and the bag-of-words representation was produced with the following parameters: $\sigma = 1$, the number of searched closest vectors is 16; these values and codebook size were evaluated in [13] and are suitable for bag-of-words creation from space-time low-level features. In the multi-kernel SVM creation process all four channels (bag-of-words representations of HOG, HOF, DT and MBH descriptors) are combined together, no searching for a better combination is performed.

The above described evaluation procedure was repeated for every class contained in the Hollywood2 dataset. For each class, we are presenting the graph of dependency between the video sample shot length and the system best accuracy, as well as the figure showing the number of frames needed to achieve 90% of state-of-the-art accuracy.

It should be noted that the first group of results (*driving car, running, fighting, eating*) corresponds well to the expectation that the accuracy will be increasing with the length of the shot. The second group (*kissing, getting out of the car, huggung, sitting up*) showed approximatelly constant accuracy depending on the length. This was probably due to the fact that the actions in these shots are recognized based on some short motions inside the actions.

The final group (*standing up, sitting down, hand shaking, answering the phone*) showed decreased accuracy depending on the length. The reason is that the actions were too short (length shown using markers in Figures 3.2 and 3.3) and so increasing the length of the shot only "increased noise" and did not bring any additional information. The expectations were also not fulfilled for generally poorly recognised actions.

Based on our experiments, for example, the running activity can be recognized in 10 frames of space-time features with 0.765 accuracy (90% of the state-of-the-art) which corresponds to the 18 frames in total and approximately 0.72s of real-time.

Chapter 4

Conclusion

The goal of this thesis was to improve human action recognition. The state-of-the-art in human action recognition has been explored, and an off-line recognition system based on multiple types of space-time features was implemented. This off-line system improves upon existing human action recognition solutions in certain situations.

The experiments presented in this thesis showed that the off-line solution was able to outperform the stat-of-the-art off-line methods for 4 of 12 action types from the Hollywood2 dataset; the results were comparable for the other 8 action types.

As the main contribution, the thesis explored a hypothesis that an optimal length of video-segments for recognition of different actions for on-line processing exists. The optimality was defined as a minimal video-segment length which provides close to off-line recognition quality (see Section 3.1). The existence of the optimal analysis length was verified experimentally by a novel algorithm which finds the optimal segment lengths.

The proposed algorithm was used to find the optimal video-segment lengths for on-line recognition for the Hollywood2 dataset with allowable quality drop set to 10%. The performed experiments showed that the optimal lengths exist for all actions in the Hollywood2 dataset and that 11 actions can be detected by using 25 video frames and only one action (fighting) requires 90 video frames. The on-line solution outperformed the off-line recognition as it was actually able to find segment lengths which gave better results than the whole videos from the dataset. The experiments showed that the actions can be divided into three basic groups: (i) the actions for which the recognition

quality increases with the segment length, (ii) the actions for which the recognition quality is nearly constant for different segment lengths, (iii) the actions for which the quality decreases with increasing segment length.

The evaluation was performed on a computer cluster due to the large amount of data and high requirements for computational resources. When further using the on-line solution, where optimal analysis length is used, the process is not so computationally demanding as above.

Future usage of results of this work includes the on-line detection in live video streams and content-based search for large video databases. In both cases the smaller detection latency and better accuracy can be achieved. The proposed verification algorithm shall be improved to be able to reach higher efficiency and smaller computational demandingness while finding the optimal analysis length. Overall, the proposed approach could be further adapted to allow automatical conversion of a general off-line recognition system to an on-line system.

Bibliography

[1] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, June 2008.

[2] O. Chapelle, P. Haffner, and V.N. Vapnik. Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5):1055–1064, Sep 1999.

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proceedings of the 14th International Conference on Computer Communications and Networks*, ICCCN '05, pages 65–72, Washington, DC, USA, 2005. IEEE Computer Society.

[4] J. Han and M. Kamber. *Data Mining, Southeast Asia Edition: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 360 Park Avenue South, New York, NY, 2006.

[5] Mihir Jain, Hervé Jégou, and Patrick Bouthemy. Better exploiting motion for better action recognition. In *CVPR - International Conference on Computer Vision and Pattern Recognition*, Portland, États-Unis, April 2013.

[6] Feng Jing, Mingjing Li, Hong-Jiang Zhang, and Bo Zhang. Support vector machines for region-based image retrieval. In *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, volume 2, pages II–21–4 vol.2, Baltimore, MD, July 2003.

[7] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *In IEEE International Conference on Computer Vision (ICCV)*, pages 432–439, Nice, France, 2003.

[8] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *In Conference on Computer Vision & Pattern Recognition*, Anchorage, Alaska, USA, 2008.

[9] Q.V. Le, W.Y. Zou, S.Y. Yeung, and A.Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3361–3368, 1420 Austin Bluffs Pkwy, Colorado Springs, CO USA, June 2011.

[10] David G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision - Volume 2*, ICCV '99, Washington, DC, USA, 1999. IEEE Computer Society.

[11] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[12] Raphael Ortiz. Freak: Fast retina keypoint. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 510–517, Washington, DC, USA, 2012. IEEE Computer Society.

[13] Ivo Reznicek and Pavel Zemcik. Action recognition using combined local features. In *Proceedings of the IADIS Computer graphics, Visulisation, Coputer Vision and Image Processing 2013*, pages 111–118, Prague, Czech Republic, 2013. IADIS.

[14] Josef Sivic and Andrew Zisserman. Efficient visual search of videos cast as text retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):591–606, April 2009.

[15] Heng Wang, A. Klaser, C. Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 3169–3176, Washington, DC, USA, 2011. IEEE Computer Society.

[16] Geert Willems, Tinne Tuytelaars, and Luc Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proceedings of the 10th European Conference on Computer Vision:*

Part II, ECCV '08, pages 650–663, Berlin, Heidelberg, 2008. Springer-Verlag.

[17] J. Wu. *Advances in K-means Clustering: A Data Mining Thinking*. Springer Theses. Springer-Verlag, Berlin, Heidelberg, 2012.

Curriculum vitae

Osobní údaje

Jméno / Příjmení	Ivo Řezníček
Adresa	Sudice 130, 680 01 Sudice
Telefon	+420776015022
Datum narození	14. prosince 1982
Mateřský jazyk	Čeština

Vzdělání

2007	Státní závěrečná zkouška - získaný titul Ing.
2005 - 2007	Vysoké učení technické v Brně, Fakulta informačních technologií, magisterský studijní program Počítačová Grafika a multimédia
2005	Státní závěrečná zkouška - získaný titul Bc.
2002 - 2005	Vysoké učení technické v Brně, Fakulta informačních technologií, bakalářský studijní program Informační technologie
2002	Maturitní zkouška
1998 - 2002	Střední odborné učiliště a Učiliště Boskovice, obor Mechanik Elektronik, nám. 9. května, Boskovice

Praxe

Experimentální činnost v oblasti zpracování obrazu a příbuzných oborů, zaměřená na způsoby zpracování příznakových vektorů.

programování v bash, perl, SQL, C, C++, matlab, powershell, .NET

Vývoj podpůrných utilit pro testování softwaru.

Automatizace testovacích protokolů za použití Powershell, .NET a TestComplete.

Vývoj informačních systémů v PHP, MySQL.

Publikační činnost

- Boháček Petr, Řezníček Ivo: Automation of testing processes, In: Proceedings of the 11th conference EEICT 2005, Brno, CZ, 2005
- Boháček Petr, Řezníček Ivo: Automation of testing processes, In: HONEYWELL EMI 2005 Proceedings of the International Interdisciplinary Student Competition and Conference, Brno, CZ, 2005
- Boháček Petr, Řezníček Ivo: Automatizace testovacích procesů, In: ACM Student Research Competition 2005, Praha, 2005
- Řezníček Ivo: Audiovisual Recording System, In: Proceedings of the 12th conference Student EEICT 2006 volume 2, Brno, CZ, 2006
- Řezníček Ivo: Object identification in image using global low-level features, In: Digital Technologies International Workshop 2008, Žilina, SK, ŽU, 2008, s. 4, ISBN 978-80-8070-953-2
- Chmelař Petr, Beran Vítězslav, Herout Adam, Hradiš Michal, Juránek Roman, Láník Aleš, Mlích Jozef, Navrátil Jan, Řezníček Ivo, Žák Pavel, Zemčík Pavel: Brno University of Technology at TRECVid 2008, In: Proceedings of TRECVID 2008, Gaithersburg, US, NIST, 2008, s. 1-16
- Beran Vítězslav, Herout Adam, Hradiš Michal, Řezníček Ivo, Zemčík Pavel: Video Summarization at Brno University of Technology, In: ACM Multimedia, New Yourk, US, ACM, 2008, s. 4, ISBN 978-1-60558-303-7
- Chmelař Petr, Beran Vítězslav, Herout Adam, Hradiš Michal, Řezníček Ivo, Zemčík Pavel: Brno University of Technology at TRECVid 2009, In: TRECVID 2009: Participant Notebook Papers and Slides, Gaithersburg, MD, US, NIST, 2009, s. 11

- Beran Vítězslav, Herout Adam, Řezníček Ivo: Video-Based Bicycle Detection in Underground Scenarios, In: Proceedings of WSCG'09, Plzeň, CZ, ZČU v Plzni, 2009, s. 4, ISBN 978-80-86943-94-7
- Řezníček Ivo, Zemčík Pavel, Herout Adam, Beran Vítězslav: SVM CLASSIFIERS CREATION IN PARALLEL CONSTRAINED ENVIRONMENT, In: Proceedings of the IADIS Computer graphics, Visulisation, Coputer Vision and Image Processing 2010, Freiburg im Breissgau, DE, IADIS, 2010, s. 535-538, ISBN 978-972-8939-22-9
- Řezníček Ivo, Bařina David: Classifier creation framework for diverse classification tasks, In: Proceedings of the DT workshop, Žilina, SK, 2010, s. 3
- HRADIŠ Michal, BERAN Vítězslav, ŘEZNÍČEK Ivo, HEROUT Adam, BAŘINA David, VLČEK Adam a ZEMČÍK Pavel. Brno University of Technology at TRECVid 2010. In: TRECVID 2010: Participant Notebook Papers and Slides. Gaithersburg, MD: National Institute of Standards and Technology, 2010, s. 11.
- ŘEZNÍČEK Ivo a ZEMČÍK Pavel. On-line human action detection using space-time interest points. In: Zborník príspevkov prezentovaných na konferencii ITAT, september 2011. Praha: Matematicko-fyzikální fakulta UK, 2011, s. 39-45. ISBN 978-80-89557-01-1.
- HRADIŠ Michal, ŘEZNÍČEK Ivo a BEHÚŇ Kamil. Brno University of Technology at MediaEval 2011 Genre Tagging Task. In: Working Notes Proceedings of the MediaEval 2011 Workshop. Pisa, Italy: CEUR-WS.org, 2011, s. 1-2. ISSN 1613-0073.
- BERAN Vítězslav, HRADIŠ Michal, OTRUSINA Lubomír a ŘEZNÍČEK Ivo. Brno University of Technology at TRECVid 2011. In: TRECVID 2011: Participant Notebook Papers and Slides. Gaithersburg, MD: United States Department of Commerce, National Institute of Standards and Technology, 2011, s. 10.
- HRADIŠ Michal, ŘEZNÍČEK Ivo a BEHÚŇ Kamil. Semantic Class Detectors in Video Genre Recognition. In: Proceedings of VISAPP 2012. Rome: SciTePress - Science and Technology Publications, 2012, s. 640-646. ISBN 978-989-8565-03-7.
- ŘEZNÍČEK Ivo a ZEMČÍK Pavel. Action recognition using combined local features. In: Proceedings of the IADIS Computer graphics, Vi-

sulisation, Coputer Vision and Image Processing 2013. Praha: IADIS, 2013, s. 111-118. ISBN 978-972-8939-89-2.

- ŘEZNÍČEK Ivo a ZEMČÍK Pavel. Human action recognition for real-time applications. In: Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods. Angers, 2014, s. 646-653. ISBN 978-989-758-018-5.