Towards Automated Assistance for Operating Home Medical Devices

Zan Gao [#1], *Marcin Detyniecki* [#2], *Ming-yu Chen* [#3], *Wen Wu* [#3],

Alexander G. Hauptmann [#3], *Howard D. Wactlar* [#3]

[1] *School of Information and Communication Engineering, BUPT, 100876, Beijing, P.R. China*

[2] *Laboratoire d'Informatique de Paris 6 - LIP6, 8, rue du Capitaine Scott 75015, Paris - France*

[3] *Computer Science Department, School of Computer Science, Carnegie Mellon University, 15213, PA, USA*

zangaonsh4522@gmail.com, Marcin.Detyniecki@lip6.fr, {mychen, wenwu, alex, wactlar}@cs.cmu.edu

***Abstract*—To detect errors when subjects operate a home medical device, we observe them with multiple cameras. We then perform action recognition with a robust approach to recognize action information based on explicitly encoding motion information. This algorithm detects interest points and encodes not only their local appearance but also explicitly models local motion. Our goal is to recognize individual human actions in the operations of a home medical device to see if the patient has correctly performed the required actions in the prescribed sequence. Using a specific infusion pump as a test case, requiring 22 operation steps from 6 action classes, our best classifier selects high likelihood action estimates from 4 available cameras, to obtain an average class recognition rate of 69%.**

***Keywords*— Multi-Camera; Human Behaviour Recognition; Medical Devices;**

I. INTRODUCTION

To help meet the increasing health care needs of our aging populations we anticipate significant increase in the use of prescribed home medical devices. To maintain their independence, individuals and/or their at-home companions must be able to prepare and apply these devices accurately and reliably. As individuals age, many are impacted by cognitive decline such that the proper sequencing of steps of a task is forgotten. Since the danger of making an error can be quite high, we are developing observational methods to detect errors in the preparation and operation of such a device, with the goal of correcting mistakes before negative side effects could occur. The goal of our project is to develop a flexible mechanism that allows a system to learn and refine representations of high-level tasks, from observation and interaction with a human operator, based on a set of underlying primitive actions that the system already understands. Examples of such home healthcare devices are respirators, oxygen pumps, ventilators and nebulizers (to help breathing), dialysis machines, infusion pumps, and apnea monitors. The primary target population for this work would be elderly people living at home, requiring support from home medical devices, with impending diminishment of cognitive function. These devices can allow a patient to live independently with minimal assistance, as long as the home medical devices provide the required health support. As a test case, we use an infusion pump which allows patients to administer intravenous medication at home.

We assume the system will be trained on video recordings of other people operating the device correctly, and later it will detect the correct operation sequence in the currently performed procedure, and, ultimately, provide corrective feedback to the user, including perhaps a portion of the recording that shows the appropriate step to be performed.

In many cases, the patient is instructed in the use of the device over multiple days by a visiting nurse, and eventually performs the operations completely unsupervised. Our work imitates this process, providing a general set of observations from other subjects used for training, and a single instance of training data from the specific subject to obtain high accuracy recognition of device operation actions.

The technical work described here consists of two components (1) training the system by recording a set of operations, (2) observing a new instance of the operation actions and recognizing what operation is being performed.

II. RELATED WORK

Many schemes have been proposed for the human action recognition. Aggarwal et al. [1] give an overview of the various tasks involved in the motion analysis of human bodies. Hu et al. [2] review the visual surveillance in dynamic scenes and analyze different research directions. Dollar et al. [3] use sparse spatiotemporal features to perform behavior recognition including human and rodent behavior. Schuldt et al. [4] construct video representations in terms of local space-time features and integrate such representations with SVM classification schemes for recognition. Laptev et al. [5] build on the idea of the Harris interest point detector and find local structures in space-time. Shechtman et al. [6] extend the notion of 2-dimensional image correlation into a 3-dimensional space-time volume, thus enabling them to correlate dynamic behaviors and actions. Liu et al. [8] use the Maximization of Mutual Information (MMI) technique to select the optimal number of words for bag-of-words algorithms for action recognition. Laptev et al. [9] address

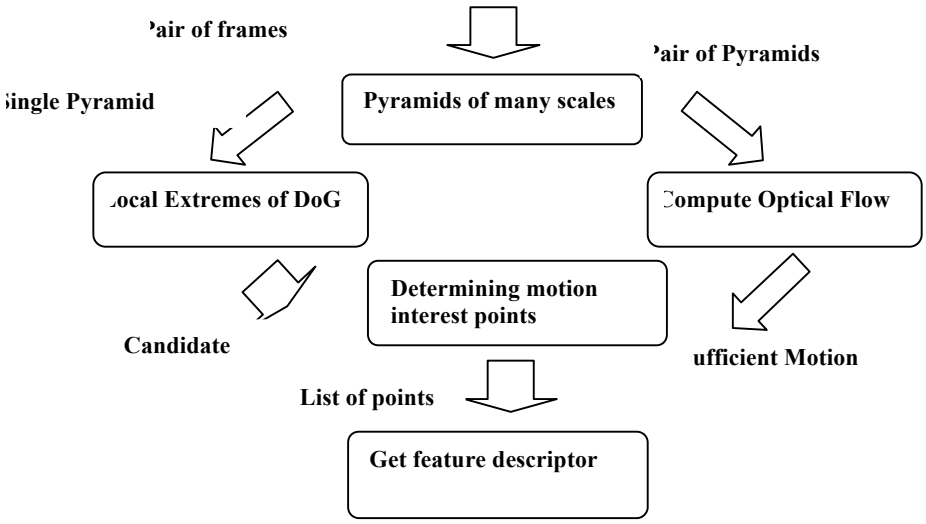

Fig 1: System flow graph of the MoSIFT algorithm. A pair of frames is the input. Local extremes of the Difference of Gaussian (DoG) images, computed by subtracting adjacent intervals, and optical flow determine the interest points for which features are described.

recognition of natural human actions with spatio-temporal interest points. Klaser et al. [13] present a local descriptor based on histograms of oriented 3D spatio-temporal gradients. Wong et al. [12] utilize the global information to yield a sparse set of interest points for motion recognition. Willems et al. [14] present the spatio-temporal interest points that are at the same time scale-invariant (both spatially and temporally). Oikonomopoulos et al. [15] detect the spatiotemporal salient points by measuring changes in the information content of pixel neighborhoods not only in space but also in time. Sun et al. [16] explore what features are suitable for different action dataset, and fuse local and holistic features to improve performance.

III. MOTION FEATURES AS BUILDING BLOCKS

Our approach has three major steps: detecting interest points, constructing a feature descriptor, and building a classifier. Detecting interest points reduces the video from a volume of pixels to compact but descriptive interest points.

This section outlines Chen's MoSIFT algorithm [19] to detect and describe spatio-temporal interest points. It was shown [19] to outperform Laptev's method [5], which also uses spatio-temporal points. MoSIFT first applies the SIFT algorithm to find visually distinctive components in the spatial domain and detects spatio-temporal interest points through (temporal) motion constraints. The motion constraint consists of a 'sufficient' amount of optical flow around the distinctive points.

A. MoSIFT Motion Interest Point Detection

The algorithm takes a pair of video frames to find spatio-temporal interest points at multiple scales. Two major computations are applied: SIFT point detection [18] and optical flow computation matching the scale of the SIFT points. Fig.1 gives the framework of the algorithm.

SIFT was designed to detect distinctive interest points in still images. The candidate points are distinctive in appearance, but they are independent of the motions in the video. For example, a cluttered background produces interest points unrelated to human actions. Clearly, only interest points with sufficient motion provide the necessary information for action recognition.

Multiple-scale optical flows are calculated according to the SIFT scales. Then, as long as the amount of movement is suitable, the candidate interest point is retained as a motion interest point.

The advantage of using optical flow, rather than video cuboids or volumes, is that it explicitly captures the magnitude and direction of a motion, instead of implicitly modeling motion through appearance change over time.

Motion interest points are scale invariant in the spatial domain. However, we do not make them scale invariant in the temporal domain. Temporal scale invariance could be achieved by calculating optical flow on multiple scales in time.

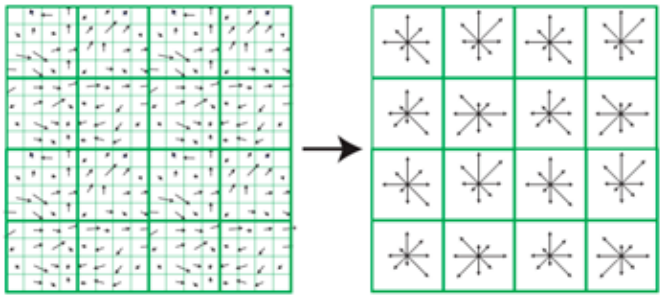

Fig.2: Grid aggregation for MoSIFT feature descriptors. Pixels in a neighborhood are grouped into 4x4 regions. As in SIFT, an orientation histogram with 8 bins is formed for each region resulting in a 128 element vector. The MoSIFT feature concatenates aggregated grids for both appearance (SIFT) and motion for a 256 element descriptors vector.

B. Motion and Appearance Feature Description

Appearance and motion information together are the essential components for an action classifier. Since an action is only represented by a set of spatio-temporal point descriptors, the descriptor features critically determine the

information available for recognition.

The MoSIFT descriptor adapts the idea of grid aggregation in SIFT to describe motions. Optical flow detects the magnitude and direction of a movement. Since, optical flow has the same properties as appearance gradient, the same aggregation can be applied to optical flow in the neighborhood of interest points to increase robustness to occlusion and deformation.

The main difference to appearance description is in the dominant orientation. For human activity recognition, rotation invariance of appearance remains important due to varying view angles and deformations. Since our videos are captured by stationary cameras, the direction of movement is an important (non-invariant) vector to help recognize an action. Therefore, our method omits adjusting for orientation invariance in the motion descriptors.

Finally, the two aggregated histograms (appearance and optical flow) are combined into the descriptor, which now has 256 dimensions. Fig.2 shows the feature description.

IV. CLASSIFICATION

For each key frame, the number of extracted key points can be different. Therefore, we use a bag-of-words (BoW) approach, based on a k-means clustering, to quantify the combined motion and appearance feature to a fixed size vector for each frame.

We aggregate all the visual words over the duration of a single event. Thus, each event is represented by a visual word histogram. We then apply a χ^2 kernel with an SVM classifier because it has been shown to be better for calculating histogram distances [17].

Whereas in text information retrieval the vocabulary size is determined by the words found in a dictionary, the size of a visual word vocabulary must be decided by the number of clusters created with a clustering process. Choosing the optimal size is a trade-off between discrimination and generalization. A small vocabulary is less discriminative since two key points may be assigned into one cluster even if they are not similar to each other. On the other hand, a large vocabulary may lack generalization power since similar key points may be assigned to different clusters. Preliminary work showed that using a moderate visual word vocabulary size leads to robust performance. Throughout our experiments, we used a 600 BoW vocabulary size.

V. EXPERIMENTAL SETUP

To study the actual pump operation procedure, we use a real Abbot Laboratories Infusion Pump (AIM Plus Ambulatory Infusion Manager). To operate this pump, the following protocol is required (simplified for readability):

1. Power on the pump device
2. Select the correct infusion program (note: this requires multiple steps) (2-5)
6. Open (un-cap) the arm port (entry tube into the body)
7. Clean the port with sterile pads
8. Open (un-cap) pump tube
9. Clean the pump tube (port) with sterile pads
10. Flush the port tube with saline syringe injections
11. Connect arm port tube to the pump tube
12. Start the infusion process on the device
13. When pump signals end of infusion, stop pump program
14. Disconnect arm port tube
15. Clean pump port tube with sterile pads
16. Cap pump port tube
17. Clean arm port with sterile pads
18. Flush the arm port tube with saline syringe injections
19. Flush arm port tube with Heparin anti-clotting agent syringe injections
20. Clean arm port tube
21. Cap arm port tube
22. Turn off pump.

At the beginning, the pump, an IV Bag, IV infusion sets, syringes and alcohol prep pads are placed on a square table. In our experiments, the saline solution and the blood-thinning agent Heparin were replaced by color-coded syringes with water and the infusion tube was not connected to the body, but the port was positioned as if it were. On the table each participant operated the pump to achieve a correct infusion. A realistic infusion protocol was designed with the help of an expert consultant. The infusion protocol had 22 steps. Although some steps can be interchanged in order, a certain critical sequence is vital. The "correct infusion steps" were demonstrated to each participant and the protocol was tried once with assistance from an experimenter acting as a trained nurse or caregiver.

As the initial data, we recorded nine subjects using the device correctly once each. We then recorded, for test purposes, the complete operation sequence again for three of the previous subjects.

The scene was recorded using 4 standard camcorders for all our 9 participants, for a total of 12 complete operations. The specifications of the recording setup are given in Fig.3. The videos were time-synchronized and each action was labeled manually to provide ground truth. Different related actions were grouped into 6 classes of behaviors: (1) Powering the pump On/Off, (2) pressing a button, (3) cleaning a port, (4) connecting/disconnecting the tubes, (5) injecting the contents of a syringe, (6) removing or attaching a cap to a tube end. Images from different camera views are shown in Fig.4.

The experiment, named leave-one-Sequence-out (LoSo), was designed as a recognition experiment: the system was given a predefined event period and was asked to determine the class of action that had taken place. We trained on all sequences from all subjects except for one sequence from one of the three test subjects. The accuracy test was performed on that held-out sequence. This experiment approximates the situation where the specific patient while practicing under supervision, is using the device and his/her particular way of moving is captured as additional personalized training data.

To compare how this patient-adapted approach compares to generic approach a second experiment, named leave-one-

Person-out (LoPo), was designed. We trained the action classifier on all the subjects, but left completely one of the three test subjects and tested the accuracy of the action classifier on the two sequences of that person. We repeated this for all three subjects with 2 sequences, and averaged the accuracy results for the same action class. This models the situation where the pump operation classifier has been trained in a lab setting by the manufacturer, but not for a specific patient.

In all cases, it was guaranteed that the classifier had never seen a particular sequence while testing on that sequence, and the same sequences (3 subjects x 2 sequences each) were to determine accuracy of the action classifiers.

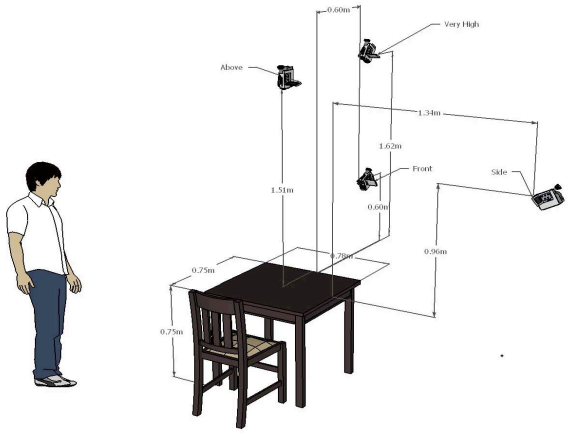

Fig 3: Setup - Exact placement of cameras for recording

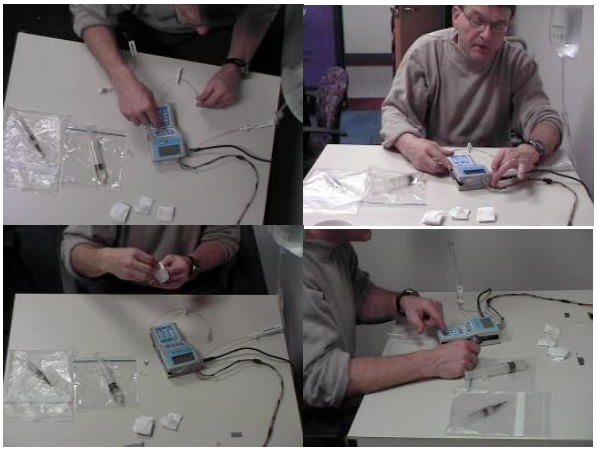

Fig 4: Example of videos from the different cameras. From left to right and top to bottom, images are from the "above" camera, the "front" camera, the "very high" camera and the "side" camera

VI. RESULTS

Table1 shows that the LoSo average action class recognition accuracy is around 61%, which is much better than a baseline random classification (17%). In addition, the train and test scheme of the patient-adapted LoSo approach improve the performance from the generic LoPo approach from 51% to 61%. What is more, the performances of all actions except for virtually random "connect/disconnect" in LoSo are much better than them in LoPo. For example, the accuracy of "power on/off" in LoPo and LoSo changes from 44% to 77%. In addition, we also see that different action classes can be detected with widely differing accuracy; the recognition per action is not close to the average with two exceptions: the well recognized "push button" action (96%) and the rarely recognized action of connecting and disconnecting (0%).

TABLE 1. ACTION CLASS RECOGNITION ACCURACY ACTION OCCURRENCE FREQUENCY, RANDOM CLASSIFICATION ACCURACY AND CLASSIFICATION ACCURACY USING THE LEAVE-ONE-PERSON-OUT (LOPO) AND LEAVE-ONE-SEQUENCE-OUT (LOSO) METHODS.

Action Class	Freq Count	Baseline (random)	Avg LoPo Accuracy	Avg LoSo Accuracy
Power On/Off	12	0.08	0.44	0.77
Push button	49	0.34	0.94	0.96
Open/CapArm/Pump	24	0.17	0.69	0.74
Flush Green/Yellow	16	0.11	0.48	0.67
Clean Arm/Pump	31	0.22	0.45	0.50
Connect/Disconnect	12	0.08	0.06	0.00
(Total)Average	144	0.17	0.51	0.61

TABLE 2: AVERAGE ACTION RECOGNITION ACCURACY FOR EACH CAMERA WITH THE LOSO SCHEME

Camera / Action	Above	Front	Side	Very High
Power On/Off	0.75	0.83	0.67	0.83
Push button	1.00	0.96	0.96	0.92
Open/Cap Arm/Pump	0.79	0.75	0.83	0.58
Flush Green/Yellow	0.75	0.56	0.88	0.50
Clean Arm/Pump	0.45	0.45	0.58	0.52
Connect/Disconnect	0.00	0.00	0.00	0.00
Average	0.62	0.59	0.65	0.56

Table 2 shows the results from each camera in the LoSo method. We observe that there is a strong variability among camera views for each action.

For instance, the "power On/Off" action is much better recognized by the front and very high camera (83%) than by the side one (67%) and the above camera (75%). In fact, this action is systematically occluded when viewed from the side camera: the button is located of the pump while the camera views it from the right.

On average, the difference between cameras is less pronounced, ranging from 59% to 65%.

The side camera is clearly the best performing view. In fact, besides the "power on/off" mentioned above, the actions "open/cap", "flush" and "clean" are much better recognized from the side camera than from any other point of view, with accuracies of 83%, 88% and 58% respectively.

In Table 3, we compared the results of a LoSo setup combining the cameras by choosing the best point of view for each action with the best camera (side view).

The average recognition of this combined-camera strategy is 69%, a moderate improvement over the best camera individually (65%).

The confusion matrix of the best camera for each action in

the LoSo scheme (Table 4) shows that the system has trouble differentiating actions "connect/disconnect" from "open/cap", which is reasonable since the movements are almost identical. In the former the subject holds two translucent tubes, while in the latter there is only one tube. The confusion is asymmetric, which can be explained by the number of examples: there are twice as many of "open/cap" than of "connect/disconnect" in every procedure sequence. The next major misclassifications result in "flushing" being confused with "cleaning" and "cleaning" confused as "open/cap". All these confusions are consistent with the perceived visual ambiguity of the actions.

TABLE 3: COMPARISON BETWEEN BEST CAMERA (*SIDE* VIEW) AND BEST CAMERA PER ACTION ACCURACY WITH LOSO

Action Class	Best Camera (*Side*)	Best Camera per Action
Power On/Off	0.67	0.83
Push button	0.96	1.00
Open/Cap Arm/Pump	0.83	0.83
Flush Green/Yellow	0.88	0.88
Clean Arm/Pump	0.58	0.58
Connect/Disconnect	0.00	0.00
Average	0.65	0.69

TABLE 4: CONFUSION MATRIX FOR THE BEST CAMERA/ACTION PAIRS

Truth \ Recognized as	Power On/Off	Push button	Open/Cap Arm/Pump	Flush Green/Yellow	Clean Arm/Pump	Connect/Disconnect
Power On/Off	0.83	0.00	0.08	0.08	0.00	0.00
Push button	0.00	1.00	0.00	0.00	0.00	0.00
Open/Cap Arm/Pump	0.04	0.04	0.83	0.00	0.08	0.00
Flush Green/Yellow	0.00	0.00	0.13	0.88	0.00	0.00
Clean Arm/Pump	0.13	0.00	0.23	0.06	0.58	0.00
Connect/ Disconnect	0.08	0.17	0.67	0.00	0.08	0.00

The next major misclassifications result in "flushing" being confused with "cleaning" and "cleaning" confused as "open/cap". All these confusions are consistent with the perceived visual ambiguity of the actions.

VII. CONCLUSION

The presented approach (motion interest points combined with $\chi 2$-kernelized SVM) provides good performance, with a general (single camera) recognition average of 61%.

The best recognition results, 69% in average action class recognition accuracy, were obtained by selecting, for each action, the best camera. Although, in terms of average performance, the use of the side camera alone provided roughly comparable results at 65% action class recognition accuracy, the observed per camera variability shows that there is potential for multi-camera combination. We are currently exploring different approaches to achieve this in a robust and effective way.

What are the implications for our broader goal? We are interested in providing assistance to patients in a home setting using a medical device without human supervision.

We believe that the high performance results, achieved on this sample device with limited training data, provide good initial evidence that automated monitoring of home medical devices may indeed be viable. In the future we will address the problem of, not only detecting action behaviors, but also locating them with respect to the steps in the protocol. To achieve the latter we are considering exploiting sequential constraints and additional sensor information from the device directly. For example, the device can know when it was turned on, and which button was pushed at what time. This would provide additional help in understanding the patient's behavior.

Further in the future, we plan to work on recognizing the actions in real-time, in order to be able to then develop live interventions to remind the patient when an action sequence was not performed correctly or in the wrong order and potentially avoid dangerous complications.

Of course, any results obtained with this pump, would also have to be generalized to other types of home medical devices, with perhaps different required actions and protocols. However we believe we have achieved the first step in demonstrating that it is feasible to apply computer vision to observe operations of a home medical device and recognize which actions are being performed.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation Grants IIS-0917072 and CNS-0751185, and by the National Institutes of Health (NIH) Grant 1RC1MH090021-0110. Zan Gao is partially supported by the NSFC (No.60772114). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the National Institutes of Health or National Science Foundation of China.

REFERENCES

[1] J. K. Aggarwal and Q. Cai. Human motion analysis: a review. IEEE Proceedings of Nonrigid and Articulated Motion Workshop, pages 90-102, 1997.

[2] W. M. Hu, T. N. Tan, L. Wang, and S. Maybank, A survey of visual surveillance of object motion and behaviors, IEEE Transactions on System, Man, and Cybernetics (T-SMC), Part C, 34(3), pages 334-352, 2004.

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pages 65- 72, 2005.

[4] Schuldt, C. Laptev, and B. I. Caputo. Recognizing human actions: a local SVM approach. ICPR(17), pages 32-36, 2004.

[5] I. Laptev and T. Lindeberg. Space-time interest points. ICCV, pages 432–439, 2003.
[6] E. Shechtman and M. Irani. Space-time behavior-based correlation - OR - How to tell if two underlying motion fields are similar without computing them?. PAMI, 29(11):2045-2056, 2007.
[7] K. Mikolajczyk and H. Uemura. Action recognition with motion-appearance vocabulary forest. CVPR, pages 1-8, 2008.
[8] J. Liu and M. Shah. Learning human actions via information maximization. CVPR, pages 1-8, 2008.
[9] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. CVPR, pages 1-8, 2008.
[10] S. Nowozin, G. o. Bakır, and K. Tsuda. Discriminative subsequence mining for action classification. ICCV, pages 1-8, 2007.
[11] A. Gilbert, J. Illingworth, and R. Bowden. Scale invariant action recognition using compound features mined from dense spatio-temporal corners. ECCV, pages 222-233, 2008.
[12] S.-F. Wong and R. Cipolla. Extracting spatiotemporal interest points using global information. ICCV, pages 1-8, 2007.
[13] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. BMVC, pages 2008.
[14] G. Willems, T. Tuytelaars, and L. V. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. ECCV, pages 650-663, 2008.
[15] A. Oikonomopoulos, L. Patras, and M. Pantic. Spatiotemporal saliency for human action recognition.ICME, pages 1-4, 2005.
[16] Sun, X., Chen, M.-Y., and Hauptmann, A. Action Recognition via Local Descriptors and Holistic Features. CVPR, Page(s): 58 - 65 June 25, 2009.
[17] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. In IJCV, 2007.
[18] David G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, pages 91-110, 2004.
[19] Chen, M-Y. and Hauptmann, A, MoSIFT: Reocgnizing Human Actions in Surveillance Videos. CMU-CS-09-161, Carnegie Mellon University, 2009.
[20] K. Schindler, and L.V. Gool. Action Snippets: How many frames does human action recognition require? In CVPR 2008.