A Comparative Study of Color and Depth Features for Hand Gesture Recognition in Naturalistic Driving Settings

Eshed Ohn-Bar and Mohan M. Trivedi[1]

***Abstract*— We are concerned with investigating efficient video representations for the purpose of hand gesture recognition in settings of naturalistic driving. In order to provide a common experimental setup for previously proposed space-time features, we study a color and depth naturalistic hand gesture benchmark. The dataset allows for evaluation of descriptors under settings of common self-occlusion and large illumination variation. A collection of simple and quick to extract spatio-temporal cues requiring no codebook encoding are proposed. Their effectiveness is validated on our dataset, as well as on the Cambridge hand gesture dataset, improving state-of-the-art. Finally, fusion of the modalities and various cues is studied.**

I. Introduction

Automatic visual interpretation of dynamic hand gestures has many potential applications in the field of human-machine interaction [1]–[5]. Hand gesture recognition is a subset of the general challenging action recognition problem, which has motivated the development of many different techniques for spatio-temporal feature extraction and analysis. Methods may employ pose estimation and tracking [6], [7], spatio-temporal templates of shape or flow using local or global patterns [8], [9], and may incorporate interest-point detection [10]–[12]. Bag-of-features techniques still stand at the forefront of the performance. Recent availability of high quality depth sensors lead to research in extending some of the aforementioned descriptors to include depth cues [13], [14]. New features specific for depth cues, such as local occupancy patterns [15], histogram of normal vectors [16], and histogram of 3D facets [17] have been proposed for the task of spatio-temporal depth-based action representation.

In this paper, an emphasis is put on performing a comparative study of several different successful spatio-temporal feature extraction methods that were previously proposed in literature, but for the purpose of hand gesture recognition from both color and depth video. The techniques and analysis are tested under visually challenging settings.

A. Contributions

Evaluation: We perform an extensive analysis of common spatio-temporal video representations. Some descriptors have been commonly evaluated on full or upper body action recognition, and need to be studied on fine detailed hand gestures. Secondly, the descriptor will be a part of a classification piepline that differs among different works. Therefore, there is a need for comprehensive study of space-time features in common experimental settings. This work aims to benchmark and motivate further research in the community by providing a comprehensive study, comparison, and evaluation methodology. We also study depth usability from a Kinect under naturalistic settings. A 'no pose required' approach is pursued for the recognition of gestures using visual spatio-temporal feature detectors and descriptors schemes [8], [9], [12], [18]–[20]. Benchmarking the feature extraction methods on the challenging dataset reveals insights into their advantages and current limitations, thereby providing an opportunity for pushing forward the performance of visual recognition systems. Runtime analysis is provided as well.

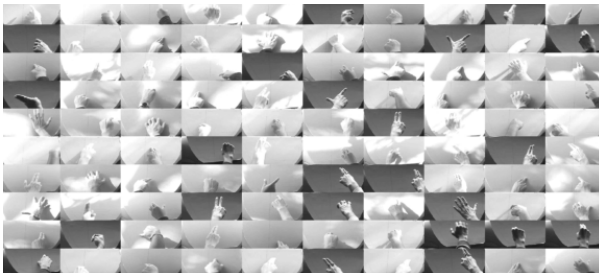


Fig. 1: Samples from the dataset studied in this paper. The middle frame of each video is visualized. The dataset allows for carrying out informative comparisons between competing approaches under visual challenges of small inter-class separation, large variation in performance of each gesture among subjects, harsh illumination changes encountered in naturalistic settings (saturation, high contrast shadows, etc.), and common self-occlusion caused by performing the gestures away from the sensor. The dataset contains both color and depth video.

Technique: We show that a combination of global low-level spatio-temporal features that can be easily and efficiently extracted stands at state-of-the-art on the Cambridge hand gesture dataset [21] and the proposed dataset in this paper. We suggest a novel feature set building on the previously proposed motion history image (MHI) extension [8] and histogram of oriented gradients (HOG) features [22]. Analysis of fusion techniques for the color and depth descriptors is also provided. The final proposed feature set is fast to extract, allowing for real-time hand gesture recognition.

II. Comparison of Space-Time Features

We benchmark several recently proposed descriptors. Although some were evaluated as color descriptors in previous literature, we benchmark descriptors on both the color and depth videos of our dataset. This, in turn, opens up the

[1]Laboratory for Intelligent and Safe Automobiles, University of California San Diego, La Jolla CA 92092, USA {eohnbar, mtrivedi}@ucsd.edu

possibly of exploiting complementary information between color and depth, especially under noisy settings where one of the modalities may provide a more reliable signal.

This work also studies a combination of features that is experimentally shown to be effective at hand gesture recognition. First, the baseline benchmarks are discussed below.

Cuboids (Dollár *et al.* [12]): As mentioned in [23], the descriptor still stands as a good benchmarking tool for space-time interest point detection. It employs a Gabor filter, followed by extracting of a 'cuboid'-a matrix of spatio-temporally windowed pixel values. The local cuboids can be processed in different methods (we use a flattened gradient) and principal component analysis (PCA) is employed to reduce dimensionality.

Harris3D [10] and **Harris3.5D** (**Ha-3D** and **Ha-3.5D**) (Hadfield and Bowden [13]): After investigation on the Hollywood3D dataset, the Harris3D (with HOG/HOF) was shown to be successful on color images compared to the Cuboids descriptor. An extension that incorporates complementary information from the appearance and depth cues by relative weighting was shown to significantly improve recognition rates (Harris 3.5D).

HOG3D (Kläser *et al.* [18]): A spatio-temporal extension of HOG, based on histograms of 3D gradient orientations. In [18] it was shown to outperform the HOG and histogram of optical flow (HOF) descriptors.

Dense Trajectories and Motion Boundary Descriptors (DTM) (Heng *et al.* [9]): Optical flow is used to extract dense trajectories, around which shape, appearance (HOG), and motion (HOF) descriptors are extracted. Finally, motion boundary histograms (MBH) are extracted along the x and y directions. This descriptor showed excellent results on a variety of action recognition datasets.

Motion History Image (MHI) (Bobick and Davis [19]): Involves computing a motion history image by successively layering image regions over time using a thresholded update rule. The pixel intensities 'decay' with time, so that initial frames are darker, and recent frames are lighter in intensity. Regions of motion are generated by differencing and thresholding to get an image D. Choosing the threshold value will have significant impact on performance as will be discussed in Section IV-A. The MHI H_τ is generated using

$$H_\tau = \begin{cases} \tau & \text{if } D(x,y,t) = 1 \\ max(0, H_\tau(x,y,t-1) - 1) & \text{otherwise.} \end{cases} \quad (1)$$

where the value of τ is the temporal length of the gesture instance. This descriptor is used to represent motion in a single static image, and can be applied to either RGB or depth input. HOG is used to produce the final descriptor for the MHI and all of the global descriptors.

HOG^2 (Ohn-Bar and Trivedi [24]): The HOG^2 (see Fig. 2) descriptor summarizes temporal characteristics in spatial HOG features extracted in each frame. The histogram descriptors are concatenated over the frames and used as an input for a second HOG application. This recently proposed descriptor is shown in this work to complement the MHI scheme. Both the HOG^2 and the MHI descriptors are extracted from the entire image frame, and not using interest points.

Descriptor	Extraction Time (in ms)	Dimensionality
GEI	0.2	128
MHI	1.9	128
HOG^2 [24]	2.8	128
EGEI (on Edge Image)	11.3	128
EMHI (on Edge Image)	13	128
DTM [9]	54	426*
Cuboids [12]	232.4	250*
HOG3D [18]	372	1000*
DSTIP [14]	774	879*
Ha-3D [10]	570	31*
HON4D [16]	40	22680
Ha-3.5D [13]	5690	234*

TABLE I: Comparison of average extraction time in milliseconds for each descriptor for *one modality* - RGB or depth. Dimensionality is shown per frame (or per interest point) for the local descriptors and per sample for the global ones. Asterisk * indicates codebook construction is needed. DSTIP is in MATLAB.

A. Depth-specific Descriptors

DSTIP (Xia and Aggarwal [14]): Spatio-temporal interest points (STIP), although successful in color video, contain noisy detections on depth video from the Kinect when flickering of the depth values occurs. The work in [14] suggests a noise suppression method, as well as a self-similarity feature to be used to describe each cuboid.

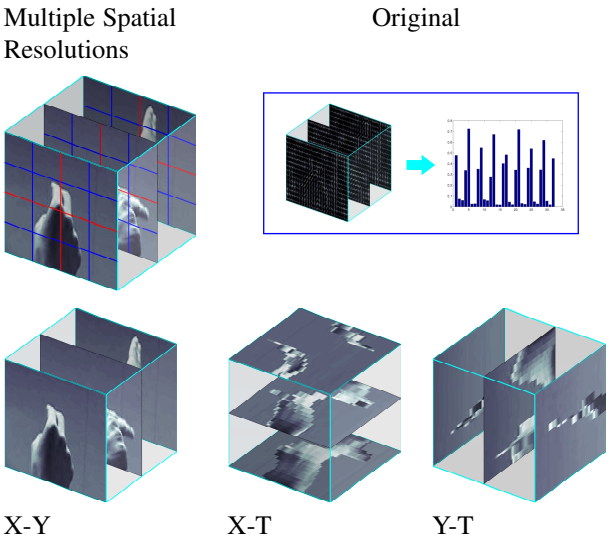
HON4D (Oreifej and Liu [16]): Incorporates a 4D histogram over depth or color, time, and spatial coordinates of the orientation of surface normals.

B. Extending MHI and HOG^2

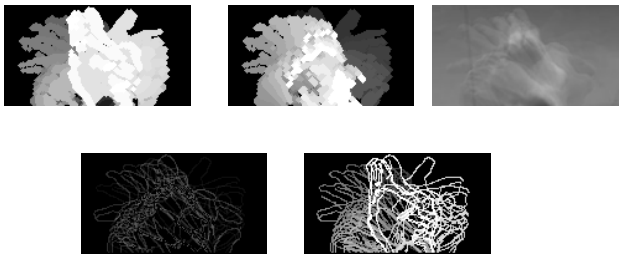
The global descriptors (MHI and HOG^2) have a significant advantage over the rest: they are small in dimensionality and fast to compute (see Table I). Therefore, training and testing repeatedly for optimizing parameters was extremely simple, unlike many of the other descriptors (e.g. simply loading the descriptors into memory in order to produce a codebook for some of the schemes is slow). Furthermore, the two descriptors combined where shown to perform well on our dataset, and so we were motivated to build upon them.

Extended HOG^2: We studied several extensions of the HOG^2 descriptor. We found that computing it at grids with multiple scales by varying the cell size (see Fig. 2) impacts accuracy favorably. Furthermore, it can be applied over orthogonal planes in the X-Y-T volume (3 in total, X-Y, X-T, Y-T, see Fig. 2). This type of temporal extension has been proposed before [25], but not with this particular descriptor.

Extended MHI: The MHI as a descriptor has been widely studied and has several advantageous properties as well as limitations. Since the only moving body in the scene is the hand, MHI is suitable for use on our dataset. However, the value of the motion threshold needs to be determined as it affects the level of noise in the MHI (see Section IV-A). The classical MHI is also sensitive to illumination changes



(a) The **extended HOG**^{2} descriptor. The original HOG^2 (shown in the blue box) can be extracted along three planes of the spatio-temporal volume (bottom row), as well as in multiple resolutions (top row-left).



(b) The **extended motion history** descriptor, incorporates a forward (top left) and backward (top middle) motion history image, as well as the gait energy image (top right). We also use edge images to compute the gait energy (bottom left) and the MHI (bottom right).

Fig. 2: Our analysis suggests that the extended HOG^2 and the extended motion history image are an effective and complementary set of spatio-temporal descriptors for state-of-the-art dynamic hand gesture recognition. The descriptors are applied both on color and depth video.

(it would be favorable to include dynamic background subtraction), occlusion and "motion overwriting" [26], while a model-based hand tracker might perform better under such settings but at a computational expense.

Recently, some of these limitations were addressed by introducing the inverse recording (**MHIINV** - MHI computed in a reverse order in order to highlight motion information from the beginning of the gesture) and the gait energy information (**GEI**) which is the temporal average of the image sequence. The work showed promising results on the ChaLearn dataset [27]. In addition to the two features, we found two extensions useful. First, a Sobel edge image is extracted at every time step. This is used for: 1) GEI on Edge Images (**EGEI**). 2) Edge Motion History Images (**EMHI**): Accumulate edges for the MHI as opposed to silhouettes (see Fig. 2). The pre-processing step of edge extraction increases robustness against background noise.

III. Evaluation Framework

Table I compares the features in terms of processing time for the feature extraction and the dimensionality (before codebook construction for the local or dense schemes). Note that the analysis is only done for RGB. Experiments were done on a desktop CPU. Generally, we follow the authors implementation and grid optimize the parameters in each method.

Codebook construction: The aforementioned methods either output a global video representation or a local sparse or dense representation. The MHI-related cues are all summarized using a HOG descriptor with 8 orientation bins (we optimize for the cell size in Section IV-C). The rest of the approaches can be encoded into a visual codebook. Although there has been some recent progress in constructing a discriminative codebook [28], k-means and Euclidean distance is still widespread. We follow the procedure in [23] where k-means is initialized 8 times and the result with the lowest error is kept. The size of the visual word codebook is determined as best on our dataset over k = 1000, 2000, 3000, 4000 (see Table I).

Classifier Choice: a support vector machine (SVM) is used in the experiments [29]. In addition to a linear SVM, we compare two other non-linear kernels. The χ^2-kernel is a common choice, defined as

$$K_{\chi^2}(X_i, X_j) = exp(-\frac{1}{2C}\sum_{k=1}^{n}\frac{(x_{ik}-x_{jk})^2}{x_{ik}+x_{jk}}) \quad (2)$$

Where C is the mean value of the χ^2 distances over all the training samples [23]. The second kernel is also common in histogram comparison, the Histogram Intersection Kernel,

$$K_{HIK}(X_i, X_j) = \sum_{k=1}^{n}\min(x_{ik}, x_{jk}) \quad (3)$$

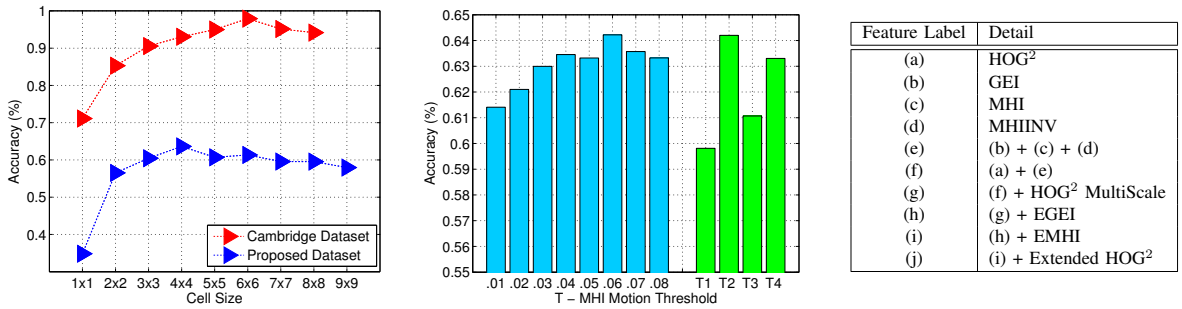
Finally, using a weighted sum of the χ^2 and HIK kernels is also reported with a slight performance increase.

IV. Experimental Evaluation and Discussion

Table II details the performance of each of the global descriptors studied in this work labeled in Fig. 3(right), as well as each of the benchmark descriptors. The top performing descriptor is highlighted for each kernel and for each modality. The reported accuracy is the average over the 8 runs of leave-one-subject-out splitting. For RGB, both the proposed feature set (Extended HOG^2 + Extended MHI) perform best together with the DTM descriptor. Interestingly, we applied the DTM to the depth video in order to see how it performs, and the results were inferior. HOG3D performed better on the depth data as opposed to color. Harris3D with HOG/HOF was shown to outperform the Cuboids detector, as in [13]. The optimal codebook size varied, for instance

RGB	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	DTM	HOG3D	Ha-3D	DSTIP	HON4D	Cuboids
Linear	27.0	15.4	24.3	29.6	38.5	42.6	45.9	47.3	48.8	**51.7**	41.3	35.8	37.2	-	-	22.1
χ^2	29.7	18.0	26.8	29.6	37.0	44.2	47.7	47.4	49.3	**49.5**	47.0	39.1	41.8	-	-	25.4
HIK	26.0	16.9	28.0	31.1	39.7	45.0	47.6	47.4	49.1	**52.2**	47.7	37.8	42.5	-	-	23.2
HIK+χ^2	26.5	16.9	27.9	31.4	39.9	45.1	47.5	47.4	49.1	**52.2**	50.1	40.4	42.0	-	-	23.4
Depth																
Linear	37.8	30.3	31.5	34.3	45.7	53.3	56.9	59.2	58.0	**60.7**	37.1	40.6	39.5	26.3	55.5	24.2
χ^2	41.9	33.8	32.8	33.8	44.6	54.3	58.1	57.5	58.9	**59.4**	40.8	43.0	40.1	29.8	57.6	25.7
HIK	38.2	32.5	30.3	36.5	47.0	55.5	56.8	58.3	59.7	**61.0**	43.2	44.2	41.4	29.4	58.3	26.1
HIK+χ^2	38.7	32.4	30.4	36.5	46.8	55.3	56.9	58.3	59.8	**61.0**	45.1	46.1	41.8	29.4	58.7	25.9

TABLE II: Performance of the different descriptors on the dataset using the two modalities, color and depth, separately. Results are shown for each of the four SVM kernels described in Section III. Feature labels for the global features are shown in Fig. 3(right). Bolded is the maximum for each kernel and for each modality.



Feature Label	Detail
(a)	HOG^2
(b)	GEI
(c)	MHI
(d)	MHIINV
(e)	(b) + (c) + (d)
(f)	(a) + (e)
(g)	(f) + HOG^2 MultiScale
(h)	(g) + EGEI
(i)	(h) + EMHI
(j)	(i) + Extended HOG^2

Fig. 3: Results for different parameter settings for the proposed feature set of extended global descriptors. **Left**: HOG cell size parameter. **Middle**: setting motion threshold to a fixed parameter (in blue) or an adaptive parameter from Eqn. 4 on our dataset. **Right**: descriptor labels for Tables II and IV.

k=1000 was shown best for HOG3D, yet k=2000 was shown best for DTM (increasing k further did not improve the results), and k=4000 for DSTIP. Combinations with DTM or HOG3D is left for future work.

Due to the their simplicity, the global descriptors allowed for extensive testing with much more ease compared to the local spatio-temporal features where just the extraction of the features over the dataset can take many hours. In order to optimize parameters, we perform the simplest fusion (concatenation) and optimize the cell size parameter (for the HOG on all the global features) and the MHI motion threshold, as shown in Fig. 3. This pushes up accuracy to 64.2%. Additional fusion techniques are discussed in Section IV-C.

A. MHI with an Adaptive Threshold

The MHI from RGB and depth requires settings a motion threshold parameter for calculating H_τ. Three adaptive techniques and one fixed technique were compared: In the fixed settings, we visually inspected the motion history image, and found settings $T_d = T_c/10$ was a good practice, where T_d is the motion threshold for the depth image, and T_c for the color image. Next we vary the lambda and evaluate the performance with all of the global descriptors from Fig. 3-left. Fig. 3-middle shows in blue the results of varying T_c using a concatenated RGBD descriptor, as well four approaches for an adaptive threshold choice detailed below.

In [8], an adaptive threshold scheme is proposed, shown in Eqn. 4 as T_1, where w and h are the width and height of the frame, N is the number of frames in the sequence,

$$\begin{array}{|l|l|} \hline T_1 = \sqrt{\frac{1}{w \times h \times N} \sum_{t=1}^{N} \sigma^2(I_t)} & T_2 = \sqrt{\frac{1}{N} \sum_{t=1}^{N} \sigma^2(I_t)} \\ \hline T_3 = \sigma(I_t) & T_4 = \sigma(D_t) \\ \hline \end{array} \quad (4)$$

and $\sigma^2(I_t)$ is the variance of the image (concatenated into a vector). We found it significantly better (with up to a 4% accuracy increase on our dataset using a concatenated RGBD descriptor and a HIK+χ^2 kernel) to use T_2. Using T_1 produced a noisy MHI due to over-sensitivity. We also experimented with using $T_3 = \sigma(I_t)$ and $T_4 = \sigma(D_t)$, where D_t is the difference image at time t, also showing better results than T_1. The results in Fig. 3 show that using a fixed threshold can slightly outperform using the adaptive thresholds considered. The results were more prominent on the Cambridge dataset, where using T_2 resulted in 93.5% as opposed to using a fixed threshold of $T = 0.04$ for all of the sequences performing at 97.9%.

B. Pre-Processing using Color Normalization

Due to the large illumination changes, we investigated several normalization approaches; one was a simple histogram equalization of the images, another was using automatic color enhancement (ACE) [30]. Pre-processing did not seem to significantly impact the performance of the proposed feature set.

C. RGBD Fusion

Generally, depth-based descriptors outperformed RGB-based ones on our dataset (Table II). We wish to study

the extent to which they are complementary. In addition, fusion techniques could leverage one modality over the other in the presence of noise. As a baseline, we use the top performing descriptor from [13], the Harris3.5D descriptor using the HOG/HOF/HODG, which showed the best results and significantly outperformed using Harris3D on RGB only. Another possible approach is to concatenate the descriptors in each frame and construct a codebook over both RGB and depth visual words, shown in Table III. Such an approach benefits the DTM descriptor, improving performance from 50.1% using RGB only to 54%. Surprisingly, this scheme produces better results than the best scheme in [13], but generally interest-point based methods did not perform well on the dataset.

We experiment with several schemes for fusion with the extended global descriptor set proposed in the paper.

Concatenation: RGB and depth features are appended for low-level, early fusion.

PCA: PCA may be used over both the RGB and depth descriptors. In Table III two possible uses of PCA are studied. First, PCA is applied over color and depth versions of each proposed global descriptor, referred to as PCA (each cue). The reduced dimension is half of the original feature vector length. Next, the different features are concatenated and jointly reduced in dimensionality to half the size.

Late Fusion: One last common fusion scheme is studied. The late fusion scheme involves learning a model for a subset of the entire RGBD vector, and then integrating the ensemble of models. This can be done by applying an operation on the probability outputs of each model (approximated using pairwise coupling [29]. Formally, let the feature vector be $\mathbf{x} \in \mathbb{R}^n$ and $\Omega = \{\omega_1, \ldots, \omega_c\}$ be the set of class labels. For an ensemble of classifiers $\{D_1, \ldots, D_n\}$, denote $d_{i,j}(\mathbf{x}) \in [0, 1]$ as the support that classifier D_i provides for the hypothesis that $\mathbf{x}$ belongs to the ω_j class. We use $d_{i,j}(\mathbf{x})$ in order to approximate the posterior probabilities for each class, $P(\omega_k|\mathbf{x})$, which is maximized to provide the final classification label. Out of a range of operations, $P(\omega_j|\mathbf{x}) = \prod_{i=1}^{n} d_{i,j}(\mathbf{x})$ was found to work well, and we compare it against learning a weighted combination with a 2^{nd}-stage SVM classifier. The experiments vary the ensemble construction. In **TwoModels**, the split is by modality ($n = 2$), such that $d_{1,j} = P(\omega_j|\mathbf{x}_{color})$ and $d_{2,j} = P(\omega_j|\mathbf{x}_{depth})$.

We also studied adding together into the late fusion different splits of the feature vector, referred to as **context**. For the TwoModels case, it involves incorporating a third model to the ensemble, $d_{3,j} = P(\omega_j|\mathbf{x}_{color}, \mathbf{x}_{depth})$.

Instead of per-modality split, we can learn a model for each of the proposed global descriptors (**ModelForEachCue**), and integrate context by fusing with models learned from concatenation of different cues. The reason why this may be useful is due to a hierarchical control over the contribution of each cue to the final posterior. This also applies to the context models, which can be used to better capture how different types of features correlate.

Table III shows the results of the experiments. Although simply training an ensemble of two models lead to a small improvement, training over individual cues and adding context leads to a significant improvement, with an accuracy of 68.1%. Fig. 4 shows the resulting confusion matrix.

Baseline	%
DTM [9]	54.0
HOG3D [18]	44.6
Ha-3.5D [13]	36.4
Proposed Feature Set	
Concatenation	64.2
PCA (each cue)	62.5
PCA (all cues)	62.3
TwoModels (color and depth)	64.5
TwoModels (color and depth) - 2^{nd} stage	64.6
TwoModels+Context	63.7
ModelForEachCue	65.6
ModelForEachCue+Context	**68.1**

TABLE III: Evaluation of different color-depth fusion techniques on our dataset. See Section IV-C for detail.

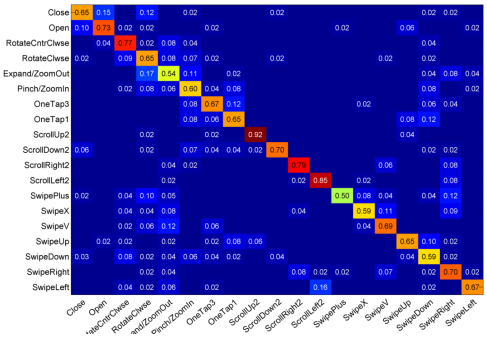


Fig. 4: Results on our dataset after color and depth fusion using ModelForEachCue+Context.

D. The Cambridge Hand Gesture

The generalization of the extended feature set proposed in the paper is tested by using the Cambridge hand gesture dataset. The dataset contains 9 dynamic gesture classes. Training is performed on the subject with normal illumination and testing is done on the other four videos with varying illumination. The extended descriptors proposed perform at an accuracy of 97.9% (Table IV).

V. Concluding Remarks

In this work we performed a comparative analysis of existing color and depth descriptors for hand gesture recognition in the car. Methodologies were also studied with different fusion schemes of the descriptors and modalities. In the future, incorporation of spatio-temporal oriented energy features could further improve recognition performance [39]. In the future, the applicability to general driver hand gestures will be studied [40], [41].

Previous Results	%	This Work	%
Baraldi *et al.* (2014) [31]	94		
Sanin *et al.* (2013) [32]	93	(j)	**97.9**
Kobayashi and Otsu (2012) [33]	92	(i)	95.8
Lui (2012) [34]	91.7	(h)	94.9
Lui *et al.* (2010) [35]	88	(g)	93.8
Harandi *et al.* (2012) [36]	86.3	(f)	92.9
Liu and Shao (2013) [37]	85	(e)	88.3
Kim *et al.* (2007) [21]	82	(d)	82.1
HMHI [37]	81	(c)	72.8
HOG/HOF [37]	79	(b)	52.7
HOG3D [37]	76	(a)	79.1
SIFT3D [37]	75		
Niebles *et al.* (2008) [38]	67		

TABLE IV: Performance using a HIK+χ^2 kernel and the descriptors in Fig. 3(right) against previous results on the Cambridge dataset.

REFERENCES

[1] F. Parada-Loira, E. Gonzalez-Agulla, and J. L. Alba-Castro, "Hand gestures to control infotainment equipment in cars," in *IEEE Intelligent Vehicles Symposium Proceedings*, 2014.

[2] V. A. Shia, Y. Gao, R. Vasudevan, K. Campbell, T. Lin, F. Borrelli, and R. Bajcsy, "Semiautonomous vehicular control using driver modeling," *TITS*, vol. 15, no. 6, pp. 2696–2709, Dec 2014.

[3] E. Ohn-Bar and M. M. Trivedi, "In-vehicle hand activity recognition using integration of regions," *IV*, 2013.

[4] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Multi-sensor system for drivers hand-gesture recognition," in *FG*, 2015.

[5] B. I. Ahmad, J. K. Murphy, P. M. Langdon, S. J. Godsill, R. Hardy, and L. Skrypchuk, "Intent inference for hand pointing gesture-based interactions in vehicles," *IEEE Trans. Cybernetics*, 2015.

[6] C. Keskin, F. Kirac, Y. E. Kara, and L. Akarun, "Real time hand pose estimation using depth sensors," in *ICCVW*, 2011.

[7] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, "Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments," *IEEE J. Sel. Topics Signal Process.*, 2012.

[8] D. Wu, F. Zhu, and L. Shao, "One shot learning gesture recognition from RGBD images," in *CVPRW*, 2013.

[9] W. Heng, A. Klser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *IJCV*, vol. 103, no. 1, pp. 60–79, 2013.

[10] I. Laptev and T. Lindeberg, "Space-time interest points," in *ICCV*, 2003.

[11] G. Willems, T. Tuytelaars, and L. V. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *ECCV*, 2008.

[12] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *VS-PETS*, October 2005.

[13] S. Hadfield and R. Bowden, "Hollywood 3D: Recognizing actions in 3D natural scenes," in *CVPR*, 2013.

[14] L. Xia and J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," 2013.

[15] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3d action recognition with random occupancy patterns," in *ECCV*, 2012.

[16] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *CVPR*, 2013.

[17] C. Zhang, X. Yang, and Y. Tian, "Histogram of 3D facets: A characteristic descriptor for hand gesture recognition," in *FG*, 2013.

[18] A. Klser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *BMVC*, 2008.

[19] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *PAMI*, vol. 23, no. 3, pp. 257–267, 2001.

[20] E. Ohn-Bar and M. Trivedi, "Joint angles similiarities and HOG^2 for action recognition," in *CVPRW*, 2013.

[21] K. Tae-Kyun, S.-F. Wong, and R. Cipolla, "Tensor canonical correlation analysis for action classification," in *CVPR*, 2007.

[22] E. Ohn-Bar and M. Trivedi, "The power is in your hands: 3D analysis of hand gestures in naturalistic video," in *CVPRW*, 2013.

[23] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC*, 2009.

[24] E. Ohn-Bar and M. M. Trivedi, "Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 6, pp. 2368–2377, Dec 2014.

[25] R. Mattivi and L. Shao, *Human action recognition using LBP-TOP as sparse spatio-temporal feature descriptor*. Berlin: Springer, 2009.

[26] M. A. R. Ahad, *Motion History Images for Action Recognition and Understanding*. London: Springer, 2013.

[27] I. Guyon, V. Athitsos, P. Jangyodsuk, B. Hamner, and H. J. Escalante, "Chalearn gesture challenge: Design and first results," in *CVPRW*, 2012.

[28] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *BMVC*, 2011.

[29] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Intell. Sys. and Tech.*, 2011.

[30] P. Getreuer, "Automatic color enhancement (ACE) and its fast implementation," in *IPOL*, 2012.

[31] L. Baraldi, F. Paci, G. Serra, L. Benini, and R. Cucchiara, "Gesture recognition in ego-centric videos using dense trajectories and hand segmentation," in *CVPRW*, 2014.

[32] A. Sanin, C. Sanderson, M. T. Harandi, and B. C. Lovell, "Spatio-temporal covariance descriptors for action and gesture recognition," in *WACV*, 2013.

[33] T. Kobayashi and N. Otsu, "Three-way auto-correlation approach to motion recognition," *Pattern Recognition Letters*, vol. 30, no. 3, pp. 212–221, 2009.

[34] Y. M. Lui, "Human gesture recognition on product manifolds," *JMLR*, vol. 13, pp. 3297–3321, 2012.

[35] Y. M. Lui, J. R. Beveridge, and K. Michael, "Action classification on product manifolds," in *CVPR*, 2010.

[36] M. T. Harandi, C. Sanderson, A. Wiliem, and B. C. Lovell, "Kernel analysis over riemannian manifolds for visual recognition of actions, pedestrians and textures," in *WACV*, 2012.

[37] L. Liu and L. Shao, "Synthesis of spatio-temporal descriptors for dynamic hand gesture recognition using genetic programming," in *FG*, 2013.

[38] J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *IJCV*, vol. 79, no. 3, pp. 299–318, 2008.

[39] K. G. Derpanis, M. Sizintsev, K. J. Cannons, and R. P. Wildes, "Action spotting and recognition based on a spatiotemporal orientation analysis," *PAMI*, vol. 35, no. 3, pp. 527–540, 2013.

[40] E. Ohn-Bar and M. M. Trivedi, "Beyond just keeping hands on the wheel: Towards visual interpretation of driver hand motion patterns," in *ITSC*, 2014.

[41] A. Fuentes, R. Fuentes, E. Cabello, C. Conde, and I. Martin, "Videosensor for the detection of unsafe driving behavior in the proximity of black spots," *Sensors*, vol. 14, no. 11, pp. 19 926–19 944, 2014.