Three-dimensional Integral Imaging for Gesture Recognition under Occlusions

V. Javier Traver, Pedro Latorre-Carmona, Eva Salvador-Balaguer, Filiberto Pla, Bahram Javidi

***Abstract*—Over the last years, three-dimensional (3D) imaging has been applied to human action and gesture recognition, usually in the form of depth maps from RGB-D sensors. An alternative which has not been explored is 3D integral imaging, aside from a recent preliminary study which shows that it can be an effective sensory modality with some advantages over the conventional monocular imaging. Since integral imaging has also been shown to be a powerful tool in other visual tasks (e.g. object reconstruction and recognition) under challenging conditions (e.g. low illumination, occlusions), and its passive long-range operation brings benefits over active close-range devices, a natural question is whether these advantages also hold for gesture recognition. Furthermore, occlusions are present in many real-world scenarios in gesture recognition, but it is an elusive problem which has scarcely been addressed. As far as we know, this work analyzes for the first time the potential of integral imaging for gesture recognition under occlusions, by comparing it to monocular imaging and to RGB-D sensory data. Empirical results corroborates the benefits of 3D integral imaging for gesture recognition, mainly under occlusions.**

***Index Terms*—Gesture recognition, integral imaging, camera array, RGB-D sensors, occlusion, classification.**

I. Introduction

Over the last decade, and due to its wide range of applications, vision-based action and gesture recognition are among the most studied topics in computer vision and machine learning [26], [2], [36], [7], [13]. Even more recent is the trend to incorporate 3D sensory due to its potential to segment (parts of) the human body and disambiguate actions [5], [8], [27].

Although literature on action recognition widely acknowledges the importance of the robustness against occlusion, it is an issue which is rarely studied in practice, and it can actually be considered an open issue [33]. Several reasons can explain this fact, such as the difficulty of addressing the problem, the elusiveness of the term and the complexity of formalizing it, and the lack of datasets [4] which include occlusions to promote a systematic study and benchmarking. Often, the occlusion considered is self-occlusion or, in the case of hand gestures, the occlusion due to the hand-manipulated object [8]. Recently, the influence of external occlusion on different known action descriptors has been studied [18].

Robustness to occlusions can be obtained by using different viewpoints or even include occlusion in training data [35]. Similarly, but out of the context of action recognition, the reconstruction of occluded objects can be dealt with the use of multiple sensors [25]. Other sensible approaches include higher-level representations, such as body parts and skeleton which can be derived from depth, usually by elaborated algorithms [27]. For instance, occluded parts can be identified and made to contribute less to the predicted action class [31].

Therefore, in general, occlusion is addressed by explicit and complex handling strategies. In sharp contrast, our work addresses the problem by exploring how the sensor itself might mitigate it, with no further assumption of the nature of the action, or occlusion, or representational issues. Certainly, this sensor-based approach is orthogonal to explicit occlusion handling strategies, but the possibility of combining them is out of the scope of this paper. It is also worth looking at the problem from the perspective of information fusion. The use of 3D Integral Imaging to fuse information from multiple sources is in essence what other approaches (multi-view cameras, RGB-D) also do, albeit differently. A novel means of computing integral images has recently been proposed [14]. Although depth information for action or gesture recognition has been exploited lately [24], [6], [22], [39], [34], [23], [29], integral imaging is not used.

To contextualize our work, it has been previously shown [30] the capabilities of integral imaging and its potential advantages and complementary properties with respect to monocular imaging for gesture recognition. Unlike RGB-D active devices such as Kinect which work in close-range indoor scenes, passive integral imaging can operate in long-range applications [21], and has shown promise to deal with challenging imaging conditions, such us turbid water [9], low illumination [10], or occlusions [16], [38], [12].

It is therefore natural to study how these methodologies perform for the particular problem of gesture recognition under occlusion, which is the main purpose of this work. Then, the main contributions of this work are:

- experiments with a new hand gesture dataset, and a study with the occlusion condition; and
- a comparison with both monocular case and another 3D imaging sensor (an off-the-shelf RGB-D device).

Throughout the paper we refer to “RGB-D” sensor as an imaging device providing color (RGB) images from a single camera plus the corresponding depth data. Although there are several technologies providing range data [5], in our experiments we will use the data from the well-known Kinect sensor [40]. On the other hand, we refer to “integral imaging” as a particular type of multi-perspective imaging using an array of cameras. In particular, we use an array of nine cameras.

All authors but the last one are with the Institute of New Imaging Technologies, Univ. Jaume I, Castelló (Spain). B. Javidi is with the Electrical & Computer Engineering Dep., Univ. of Connecticut, Storrs, USA. (e-mails: {vtraver,latorre,salvadoe,pla}@uji.es, bahram@engr.uconn.edu)

II. Methodology

The three methodological aspects of the work are the computation of integral images, the gesture representation and recognition, and a depth-based mechanism to mimic the filtering effect of integral imaging on a conventional RGB-D sensor. These aspects are introduced subsequently.

Integral Imaging. Synthetic Aperture Integral Imaging (SAII) is an auto-stereoscopic technique based on the use of an array of cameras (or one moving camera) to acquire a series of images (the *elemental* images) of a scene from slightly different perspectives. Since each of the elemental images provides a different view of the 3D scene, the 3D scene can be reconstructed using a computer synthesized virtual pinhole array for inverse mapping of each elemental image into the so-called object space [37]. To that end, the elemental images are computationally overlapped according to [15]:

$$I(x,y,z) = \frac{1}{O(x,y)} \sum_{r=0}^{R-1} \sum_{c=0}^{C-1} E_{rc}\left(x', y'\right), \qquad (1)$$

with

$$x' = x - r \cdot \frac{N_x \cdot p_x}{s_x \cdot M}, \qquad y' = y - c \cdot \frac{N_y \cdot p_y}{s_y \cdot M}.$$

where $I(x,y,z)$ is the reconstructed $3D$ image intensity at depth z, that will be referred to as the integral image; x and y are the indexes associated to each pixel position; E_{rc} is the intensity of the elemental image acquired by the camera at the r-th row and c-th column in the array; $N_x \times N_y$ are the dimensions of the images (in pixels); p_x (p_y) is the horizontal (vertical) pitch (mm) between neighbouring cameras; M is the magnification factor; $s_x \times s_y$ is the physical size of each camera sensor; and $O(x,y)$ is the overlapping number matrix, representing the number of cameras contributing to each (x,y) position.

Gesture recognition. A standard Bag of visual Words (BoW) built from local spatio-temporal interest points (STIPs) [19] was followed, with the main steps being:

Interest points detection: The STIP detector available at [20] was used, and all detected STIPs were kept by using a zero threshold in the function evaluating large spatio-temporal variations.

Local descriptors and their quantization: In turn, these STIPs are locally characterized by spatial gradients and optical flow. The resulting descriptors were vector quantized by unsupervised clustering by using k-means [17].

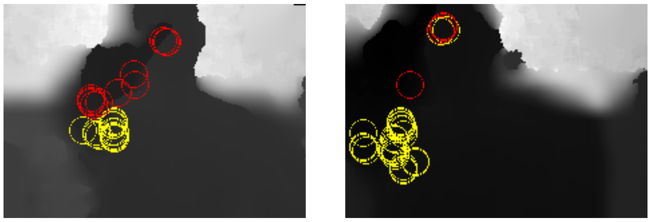
Video representation: Since a different number of STIPs are detected in each video, a histogram of words is computed per video by counting the memberships of the descriptors to the existing clusters. Therefore, each video is represented by a fixed-length feature vector (a K-bin histogram).

Recognition by classification: Finally, any standard supervised learning scheme can be used for gesture recognition by using the histogram representation of the videos for both training a classifier, and evaluating its predictive performance on unseen gestures.

Further details of these steps can be found elsewhere [30].

```
Input: Depth map $D(x,y,t)$, and
set $P$ of STIPs detected on gray-level video
Output: Set $Q$ of selected STIPs from $P$
$Q \leftarrow \varnothing$;
foreach $p(x,y,t) \in P$ do
    $S \leftarrow$ FindNeighbourSTIPs($P,p;R$);
    $S' \leftarrow$ SelectSTIPsByDepth($S,D,t;\mathcal{Z}$);
    if $|S'| \geq M$ then
        $Q \leftarrow Q \cup \{p\}$
    end
end
return $Q$
```

Algorithm 1: Depth-based STIP selection



(a) about half are kept (15 / 29) (b) almost all are removed (32 / 37)

Fig. 1. Examples of STIPs kept (red) and removed (yellow) by DBF. STIP locations are drawn on the depth maps, but they are actually detected on the RGB videos

Depth-based filtering in RGB-D. The integral images used were those reconstructed at a depth where the hand was subjectively judged to be mostly at focus. Therefore, for comparison purposes, we apply a procedure that has a similar effect in RGB-D sensors and makes use of the same "oracle" (prior knowledge of hand depths). In particular, STIPs were detected on the RGB images and then filtered (kept or removed) based on the corresponding depth. The intuitive idea is to keep STIPs that are surrounded by others at similar depths and close to depth hand. More concretely, a STIP at location (x,y) is kept if and only if at least M other STIPs at frame t are found within a square of size $2R$ centered at (x,y) that have a depth value within a given set $\mathcal{Z}$ of depth values. Algorithm 1 formalizes this notion.

The values for the minimum number of points M and the maximum neighbourhood size R were heuristically set as $M = 15$ and $R = 10$, and the set of allowed depths was the interval $\mathcal{Z} = [d - 100, d + 100]$, with d being the depth value selected interactively from the depth maps. Depth units are millimeters and distance units are pixels. A hole filling procedure was applied to the depth maps with a cross-bilateral filter [28].

In general, it is noticeable the reduction in the number of STIPs due to the depth-based filtering (DBF). At least in some cases it is observed that most STIPs are removed in videos with more noisy STIPs (i.e. those found at the face or in parts other than the hand). This suggests that noisy STIPs are generally filtered out, and therefore this DBF can potentially be helpful in better characterizing gestures. The effect of DBF is illustrated in Fig. 1; typical STIPs that can be noisy and are filtered out correspond to those in the forearm (Fig. 1a).

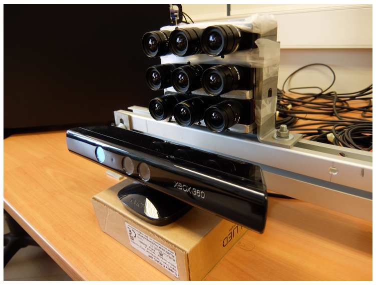


Fig. 2. The two sensors used in the experiments: the 3 × 3 camera array on top and the Kinect below

III. Experiments and results

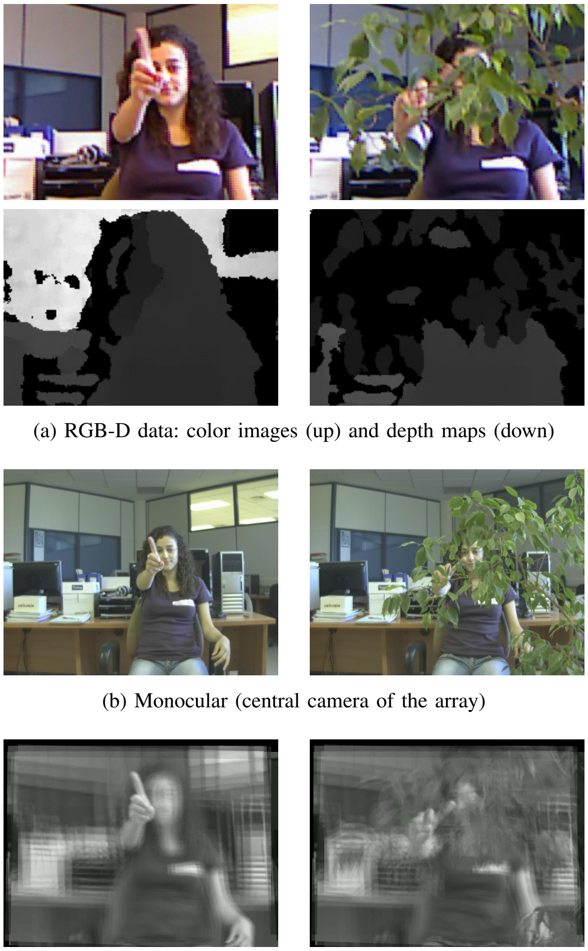
Setup. Two 3D imaging methodologies are used and compared. On the one hand, integral images were generated by synthetic aperture integral imaging from a 3×3 camera array. On the other hand, for the RGB-D data, the popular Kinect device was used. Eleven subjects were asked to perform three different gestures twice in front of the camera array, both with an unoccluded view and with the occlusion of a plant. The Kinect was placed just under the camera array so that the gestures were recorded from a very similar viewpoint (Fig. 2). The same gesture was recorded at the same time by the nine array cameras as well as by Kinect. A detailed description of the dataset is available at [1].

Kinect's images have larger field-of-view (FOV) but less resolution (640×480) than the images of the cameras of the array (1024×768) and, in turn, than the integral images. Therefore, in order to perform a fair comparison, cropping and resizing were done accordingly to have comparable effective resolutions of the region of interest (subjects' upper bodies).

Sample images (Fig. 3) illustrate the visual data under RGB-D, monocular, and integral imaging in unoccluded and occluded views. It is worth noticing the significant amount of occlusion of the plant, how noisy the depth map is under occlusion, and how integral images manages to "smooth out" much of the occluding leaves by properly focusing at the hand's depth by means of the synthetic aperture reconstruction process. The videos taken by the central camera of the array are used for the monocular condition.

Three different local visual descriptors were tested: histogram of gradients (HOG), histogram of optic flow (HOF) and their concatenation (HOG+HOF). Recognition performance was similar in the three cases, and therefore only the performance with one of them (HOG+HOF) is reported. Histograms were L_1-normalized, then individual features independently re-scaled to the range $[0, 1]$, and finally the histograms were L_2 normalized. Since different performances can be expected from visual vocabularies of different sizes, but there is no clear guideline of which size is most appropriate in which condition, then a range of vocabulary sizes $K \in \{10, 25, 50, 100, 200, 500, 1000, 2000\}$ was tested. The k-means implementation of the VLFeat library [32] was used.

For classification, two Support Vector Machines (SVMs) [11] were tested: a linear one, and a non-linear one with a Radial-Basis Function (RBF) kernel. Since similar performance was observed in both cases, only the performance with the linear SVM is reported. The LIBSVM [3] implementation of SVM was used. The parameter C in SVMs was chosen from the set $\{10^e : e \in \{-4, -3, \ldots, 4\}\}$ by cross validation. To estimate gesture performance, a "leave-one-subject-out" protocol was used. Additionally, given the random nature of k-means, the entire process (clustering + learning + classification) was repeated $n = 10$ times and the average accuracy reported. The performance plots include these averages and their standard errors as a measure of variance. In some cases, it may occur that the number of data points is less than the size K of the vocabulary, and therefore the clustering cannot be carried out for that particular size and larger ones.



(a) RGB-D data: color images (up) and depth maps (down)

(b) Monocular (central camera of the array)

(c) Integral images (hand-depth plane)

Fig. 3. Illustrative images of (a) RGB-D data, (b) monocular, and (c) integral imaging without occlusion (left) and with occlusion (right).

In the occlusion case, only the STIPs from videos of the non-occluded gestures are used at training time, since in practice one usually has only "clean" gestures for training, and occlusions happen unpredictably at test time. In other words, using occlusions at training time would imply we know in advance which particular kind of occlusion will happen and in which context, but this rarely happens in practical settings.

Effect of DBF. To study the effectiveness of the DBF procedure, we compare the performance of using the whole set of STIPs detected in the RGB Kinect's images with the filtered set of STIPs resulting from the DBF. It was found that DBF is effective as long as the vocabulary is large enough ($K > 500$). For instance, the accuracy was about 2 percent points higher than with RGB (for $K = 1000$). In fact, we tested with larger K ($K \in \{3000, \ldots, 7000\}$) so as to find out what happened with even larger vocabularies. A steady better performance of DBF was observed, even with more remarkable differences of about 8 percent points for $K = 4000$ and $K = 5000$. However, since less number of STIPs results from DBF, not enough data points were available to use $K \geq 6000$ clusters. The performance trend was however clear with the tested K.

Comparing the three sensory modalities. When comparing the three sensory modalities with and without occlusion (Fig. 4) the following observations can be made:

Occlusion: All modalities are very sensitive to the occlusion condition, which is also an indication of its severity. A drop of about 20 percent points occurs in all cases. In general, compared to the no-occlusion case, larger vocabularies are required under occlusion to get better performance.

II vs Monocular: As expected, better performance is achieved with II than with monocular images. The difference is more noticeable under occlusion, and the performance gap roughly increases with the vocabulary size. This clearly suggests the superiority of integral images to cope with occlusions.

II vs RGB-D: Without occlusion, integral imaging is more effective than RGB-D for small vocabularies, but RGB-D (through the DBF mechanism) outperforms II for larger vocabularies. It is interesting to note how II achieves very good performance even for the smallest vocabularies, which suggests these visual words are more expressive, and lend themselves to more efficient computations and less memory requirements. However, although the DBF has some positive effect without occlusions (as also discussed above), it is not sufficient to deal with occlusions, where integral imaging is clearly a better option. To understand the reasons behind this different performance between II and RGB-D, it might be good to remind what each of them is performing: in RGB-D, DBF removes some potentially noisy STIPs detected at monocular (RGB) images, whereas the STIPs detected from the integral imaging are different from the monocular case.

RGB-D vs Monocular: Without occlusion, RGB-D outperforms monocular imaging, but under occlusion DBF tends to work worse than monocular. This may be due to the fact that STIPs are removed with the DBF procedure, and this may filter out some "good" as well as "bad" STIPs.

It is important to note that no explicit occlusion-handling strategy is used; integral imaging deals naturally with occlusion, as a built-in feature resulting from its focusing ability.

It can be noticed that very few words (just 10) suffice to have reasonable and steady performance with Monocular and II. For II, we checked with $K < 10$ to find out the minimum vocabulary size, and performance drops to $\approx 65\%$ with $K = 5$. Therefore, $K \approx 10$ seems the minimum required number of words.

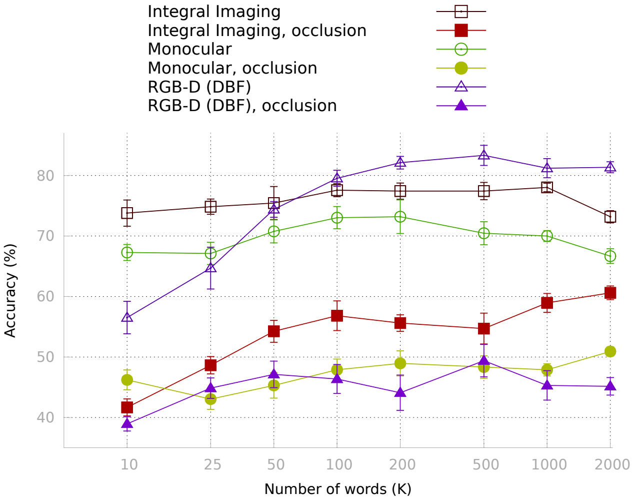


Fig. 4. Comparing Monocular, RGB-D and Integral imaging, in the three cases both without and with occlusion. The horizontal axis is shown in logarithmic scale to better accommodate the wide range of vocabulary sizes.

TABLE I
CHANGE IN AVG. ACCURACY (%) WITH RESPECT TO THE LOW RESOLUTION CASE, IN MONOCULAR AND II (* = OCCLUSION)

K ▷	10	25	50	100	200	500	1000	2000
Mono	-20.8	0.0	+2.7	+2.3	+4.4	+7.0	+6.8	+6.5
II	-12.9	-1.8	+6.2	+3.6	+4.7	+5.6	+5.2	+7.1
Mono*	-12.6	-5.5	-4.2	-6.2	-3.9	+2.8	+9.7	+6.3
II*	+3.3	+3.0	+0.3	-0.8	+4.1	+10.0	+2.4	-0.5

Resolution issue (Table I). Without occlusion, a higher spatial resolution benefits similarly Monocular and II, with an accuracy increase of about 5 percent points. At higher resolution, bigger vocabularies are required to get a steady performance, possibly because more STIPs are found. Under occlusion and in monocular, performance improves with resolution at larger vocabularies ($K \geq 500$). In general, these results can be interpreted as that the performance decays with resolution more clearly in Monocular case than in II, a sign that II can rely on its focusing ability besides the resolution quality.

IV. Conclusions

Experimental results suggest that passive 3D integral imaging offers advantages over monocular imaging even with the presence of occlusions. Without occlusion, integral imaging behaves roughly "on par" with RGB-D with the simple depth-based STIP filtering mechanism. In some cases (e.g. large visual vocabularies), RGB-D is even more effective than integral imaging. However, under occlusion, integral imaging outperforms RGB-D.

It has also been observed that good spatial resolution is much more important in monocular images than in integral images since the latter can additionally rely on a good "focusing" operation.

Despite the fact that the dataset collected and used is small, it is representative enough to have a prospective assessment of integral imaging capabilities in relation to other 3D sensor modalities, specially for the case when there are occlusions.

Acknowledgments. This work has been partly supported by Generalitat Valenciana (through grant PROMETEOII2014062).

References

[1] Hand gesture dataset for testing integral imaging under occlusions. http://www.vision.uji.es/II_hand_gestures, Nov. 2016. Computer Vision Group, Institute of New Imaging Technologies, Jaume-I University.

[2] J. Aggarwal and M. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):16:1–16:43, Apr. 2011.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[4] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Comp. Vision & Image Understanding*, 117(6):633–659, 2013.

[5] L. Chen, H. Wei, and J. Ferryman. A survey of human motion analysis using depth imagery. *Pattern Recognition Letters*, 34(15):1995–2006, 2013.

[6] L. Chen, H. Wei, and J. Ferryman. Readingact RGB-D action dataset and human action recognition from local features. *Pattern Recognition Letters*, 50:159–169, 2014.

[7] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles. Advances in human action recognition: A survey. *CoRR*, abs/1501.05964, 2015.

[8] H. Cheng, L. Yang, and Z. Liu. A survey on 3D hand gesture recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, PP(99):1–1, 2015.

[9] M. Cho and B. Javidi. Three-dimensional visualization of objects in turbid water using integral imaging. *Journal of Display Technology*, 6(10):544–547, Oct. 2010.

[10] M. Cho, A. Mahalanobis, and B. Javidi. 3D passive photon counting automatic target recognition using advanced correlation filters. *Optics Letters*, 36(6):861–863, Mar. 2011.

[11] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[12] M. Ghaneizad, H. Aghajan, and Z. Kavehvash. Three-dimensional reconstruction of heavily occluded pedestrians using integral imaging. In *Proc. of the 10th Intl. Conf. on Distributed Smart Camera*, pages 1–7. ACM, 2016.

[13] S. Herath, M. T. Harandi, and F. Porikli. Going deeper into action recognition: A survey. *CoRR*, abs/1605.04988, 2016.

[14] S. Hong, D. Shin, B. G. Lee, A. Dorado, G. Saavedra, and M. Martínez-Corral. Towards 3D television through fusion of Kinect and integral-imaging concepts. *Journal of Display Technology*, 11(11):894–899, Nov. 2015.

[15] S.-H. Hong, J.-S. Jang, and B. Javidi. Three-dimensional volumetric object reconstruction using computational integral imaging. *Optics Express*, 12(3):483–491, Feb. 2004.

[16] S.-H. Hong and B. Javidi. Distortion-tolerant 3D recognition of occluded object using computational integral imaging. *Optics Express*, 14(25):12085–12095, Dec. 2006.

[17] A. K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recogn. Lett.*, 31(8):651–666, June 2010.

[18] I. Jargalsaikhan, C. Direkoglu, S. Little, and N. E. O'Connor. An evaluation of local action descriptors for human action classification in the presence of occlusion. In *International Conference on Multimedia Modeling*, pages 56–67. Springer International Publishing, 2014.

[19] I. Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64(2–3):107–123, Sept. 2005.

[20] I. Laptev. Space-time interest points (STIP), 2011. Software available at http://www.di.ens.fr/~laptev/download.html.

[21] D. LeMaster, B. Karch, and B. Javidi. Mid-wave infrared 3D integral imaging at long range. *J. Display Technol.*, 9(7):545–551, July 2013.

[22] C. Liang, E. Chen, L. Qi, and L. Guan. Improving action recognition using collaborative representation of local depth map feature. *IEEE Signal Processing Letters*, 23(9):1241–1245, Sept. 2016.

[23] A.-A. Liu, W.-Z. Nie, Y.-T. Su, L. Ma, T. Hao, and Z.-X. Yang. Coupled hidden conditional random fields for RGB-D human action recognition. *Signal Processing*, 112:74–82, 2015.

[24] C. Lu, J. Jia, and C. K. Tang. Range-sample depth feature for action recognition. In *Proc. IEEE Comp. Vision and Pattern Recog. (CVPR)*, pages 772–779, June 2014.

[25] T. Nasrin, F. Yi, S. Das, and I. Moon. Partially occluded object reconstruction using multiple Kinect sensors. In *Proc. SPIE*, volume 9117, pages 91171G–91171G–5, 2014.

[26] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010.

[27] L. L. Presti and M. L. Cascia. 3D skeleton-based human action classification: A survey. *Pattern Recognition*, 53:130–147, 2016.

[28] N. Silberman. NYU Depth Dataset V2. http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html.

[29] Y. Song, S. Liu, and J. Tang. Describing trajectory of surface patch for human action recognition on RGB and depth videos. *IEEE Signal Processing Letters*, 22(4):426–429, Apr. 2015.

[30] V. J. Traver, P. Latorre-Carmona, E. Salvador-Balaguer, F. Pla, and B. Javidi. Human gesture recognition using three-dimensional integral imaging. *J. Opt. Soc. Am. A*, 31(10):2312–2320, Oct 2014.

[31] J. S. Tsai, Y. P. Hsu, C. Liu, and L. C. Fu. An efficient part-based approach to action recognition from rgb-d video with bow-pyramid representation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2234–2239, Nov. 2013.

[32] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.

[33] M. Vrigkas, C. Nikou, and I. Kakadiaris. A review of human activity recognition methods. *Frontiers in Robotics and AI*, 2(28), 2015.

[34] J. Wan, G. Guo, and S. Z. Li. Explore efficient local features from RGB-D data for one-shot learning gesture recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 38(8):1626–1639, Aug. 2016.

[35] D. Weinland, M. Özuysal, and P. Fua. Making action recognition robust to occlusions and viewpoint changes. In *European Conf. on Computer Vision*, pages 635–648, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[36] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Comp. Vision & Image Understanding*, 115(2):224–241, 2011.

[37] X. Xiao, B. Javidi, M. Martínez-Corral, and A. Stern. Advances in three-dimensional integral imaging: sensing, display, and applications. *Applied Optics*, 52(4):546–560, 2013.

[38] T. Yang, W. Ma, S. Wang, J. Li, J. Yu, and Y. Zhang. Kinect based real-time synthetic aperture imaging through occlusion. *Multimedia Tools and Applications*, 75(12):6925–6943, 2016.

[39] M. Yu, L. Liu, and L. Shao. Structure-preserving binary representations for RGB-D action recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 38(8):1651–1664, Aug. 2016.

[40] Z. Zhang. Microsoft kinect sensor and its effect. *IEEE MultiMedia*, 19(2):4–10, Feb. 2012.