RSS2011
Robotics Science and Systems 2011
June 27-30, 2011
Los Angeles, CA
USA

Reviews For Paper

Paper ID	110
Title	Voxelized Shape and Color Histograms for RGB-D

Masked Reviewer ID: Assigned_Reviewer_1

Review:

Question	
Please provide the technical review summary (write 1-3 full thorough paragraphs on technical	The authors propose a descriptor combining color and shape information. It is based on earlier feature detectors that separately address color and shape; the authors make several modifications and then concatenate the two features together. One of the interesting ideas in the paper is to compute a descriptor that is linear, in the sense that the descriptor for a large region is the sum of the descriptor of its parts. This allows a variety of computational optimizations. The authors begin with an overview of the earlier descriptors upon which their work is based: ColorCHLAC and GRSD. The presentation of both of these sections could be clearer and more self-contained. The summations in eqns 2-4 are over an unspecified voxel grid, for example. The handling of missing data in eqn 1-- by assigning a zero vector-- seems odd. It is not obvious that this is the "right thing to do". (Why not, for example, sum only over the voxels which exist?) Several other assumptions are silently made; evidently, we are using a 3x3x3 grid (yielding 26 neighbors), but the reader is left to figure this out on their own. Seemingly important details (such as the handling of binarized colors) are not described at all. The omission of these details makes the paper harder to read, as the reader must jump back and forth between different papers seeking out the relevant information. Similarly, other "magic numbers" suddenly appear: we end up with 12 or 21 elements without any explanation. These shortcomings make replication needlessly difficult! GRSD is extremely tersely described. The original version involves ray tracing, but the authors claim that it is similar to ColorCHLAC (which is essentially computing moments over the voxel data). It would be impossible to reconstruct the author's algorithm--- even approximately--- from this poorly documented section. The authors claim that their modified approach is two orders of magnitude faster, but don't characterize the tradeoffs that result from their approximations. In section V.A., the authors claim that their method provides a method for classifying partially visible objects. This claim is unsupported. In the synthetic data, the authors perform testing with artificial noise spanning from 0.5mm to 5.0mm. The reader has no way to know whether this is a representative range. What is the real-world noise of the Kinect, for example?

contributions, strengths, and weaknesses.	Figure 4 shows an anomolous cross-over point at 3.5mm noise, where the relative performance of the algorithms seems to invert. The paper doesn't even comment on this interesting result! Section VI.C. claims that training from views similar to that obtained by a real robot is an important aspect of their work. This claim is unsupported. (It being a "reasonable sounding" statement isn't enough!) Table II requires a better visualization. The parsing and interpreting the significance of the 160 numbers in this table is too difficult. What am I looking for? What are the interesting "take home" messages? The authors state (VI.C.3) that the results for VOSCH were "slightly less convincing". What does this mean? It's also interesting to note that SVM is *dominating* LSM. More analysis would be welcome here. The final experiment, at the end of section VI, involves an unspecified object segmentation algorithm, unspecified objects whose size is evidently known in advance, and uses hand-tuned (and unspecified) parameters. It is totally uncontrolled, unreplicatable, and thus virtually useless from an evaluation perspective. Overall, the authors present another descriptor that is cobbled together from earlier descriptors. They present some evidence that the descriptor "kinda works", but there's no meta-analysis. What do the author's experiments tell us about how to describe shape and color? Do these suggest directions to go in the future? The author's specific implementation seems ad-hoc, and the design decisions are not supported by any data. (Why CHLAC instead of some other feature? Why GRSD instead of some other feature? Why concatenation, as opposed to some other combination? How should interest points be selected? And so on.) Imprecise statements and generally unclear writing are significant weaknesses of the paper which make it unsuitable, in my opinion, for RSS.
Please provide the paper's structural or text errors (example: Figure 4 caption says z=x+y but Section II, paragraph 3 says z=x+a.).	Authors should check that acronyms are expanded before first use, and help the reader keep straight the many acronyms which appear throughout the paper. Most of these are not particularly suggestive of what they mean. VI.C.3. "It has however the advantage of exploiting...". The antecedent for it is unclear.

Masked Reviewer ID: Assigned_Reviewer_2

Review:

Question	
	The authors present two new descriptors (VOSCH and ConVOSCH) based on shape and color histograms which shall be used for object recognition. Thereby, the author assume accuarte 3D point clouds registered RGB

Please provide the technical review summary (write 1-3 full thorough paragraphs on technical contributions, strengths, and weaknesses.	information -- obtained using a RGB-D camera (=KINECT) -- which is transformed into a voxel represenatation. The two desciptors are based on two previously pusblished works for 3D shape description (GRSD) and 3D color description (ColorCHLAC.) ConVOSCH is simply a combination (concatination+normalizing) of GRSD and ColorCHLAC. However, VOSCH modifies the original descriptors in order to be rotation invariant and additive - sacrificing some accuracy. For object recognition, first a classifier (LSM or SVM) are trained on a test set of objects. Then, detection runs in real-time in a fraction of a second per object. A simulation experiment was performed on a set of 49 generic objects (7 color, 7 shapes, all combinations). Since ColorCHLAC has problems with distingushing shapess, and GRSD is totally ignorant to colors, the resuls where as expected (VOSCH>ConVOSCH>ColorCHLAC>GRSD) and demonstrate the usefulness of combinining color and shape. While the method seems to works well in theory (supported by the simulation experiments), the real image/real objects experiments are less convincing: ConVOSCH(76.4%) > ColorCHLAC(74.6%) > VOSCH(70.4%) > GRSD(24.5%) Especially the fact that VOSCH -- which was best in the simulation experiment and also seems to be conceptually superior because of its rotational invariance -- is superseded by ColorCHLAC is surprising and requires further explanation. Also, a pipeline for detection of objects in clutter using intral feature tables are presented -- however only very preliminary qualitative results are shown. To summarize, a new methods for object detection using color and shape is presented and is nicely supported by a set of simulation experiments. However the real-object experiment is less convincing.
Please provide the paper's structural or text errors (example: Figure 4 caption says z=x+y but Section II, paragraph 3 says z=x+a.).	Large portion of Section VI.D are not Expriments/Results and should be presented in seperate section. IV a: "f(x) \in R^6" --> "f(x) \in \mathbb{N}^6" "(36*13)" --> "(36\cdot 13) "

Additional comments to author(s)	Is ConVOSCH rotation invariant or not? From the description I'd assume it's not because it is based on COLORCHLAC. However, in Sec IVD it says: "but on the other side makes the feature to the extent rotation-invariant" The authors do not comment much on the main results in TABLE IV. However, an indepth discussion and interpretation would be very relevant here. In Eq. 2-4: What is the "range" of the sums? There is no index below the sum symbols

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question	
Please provide the technical review summary (write 1-3 full thorough paragraphs on technical contributions, strengths, and weaknesses.	The paper proposes the use of voxel-based color-plus-shape feature descriptors for 3-d object classification. The main contribution is the introduction of a histogram feature that combines an existing voxel-based color-histogram feature (ColorCHLAC) and an existing non-voxel-based shape-histogram feature (GRSD) Two versions of this feature are introduced: rotation-variant (ConVOSCH) and rotation-invariant (VOSCH). One strength of the approach is its speed of feature computation. Other claimed advantages of the new feature over existing ones are robustness to occlusion and the combination of color and shape in a unified manner. The final contribution of the paper is a 63-object database of RGB-D views. They compare their novel descriptors to the 2 existing inspirational descriptors on synthetic and real data. I am aware that the runtimes quoted in the paper compare favorably to those obtained with point-based features on very dense point clouds, but a numerical comparison isn't given (beyond stating that voxel-based GRSD is 2 orders of magnitude faster than original point-based GRSD). A major weakness is the results on real data, which show that ConVOSCH and ColorCHLAC both outperformed the rotationally invariant VOSCH descriptor, and the increased accuracy of ConVOSCH over ColorCHLAC is minimal. A second weakness is that, though using the linear subspace method to locate objects in cluttered scenes is one of the stated advantages of the proposed method, the demonstration of this is limited to three images of "correctly detected objects", clearly insufficient justification. The experimental results are convincing for easy synthetic data. Confusingly, the authors state that "the objects were uniformly colored and symmetrical" so the rotational variance "did not expose in the table"; on the contrary, the objects are not _rotationally_ symmetrical and it seems that random rotations did significantly lower the results of the rotationally-variant features. The best results were obtained with the SVM, so I am also not convinced that the LSM classifier is helpful. To summarize, the key limitations of this paper are:

	- limited novelty over existing descriptors - virtually no improvements over existing descriptors on real data. Additionally, LSM is significantly outperformed by SVM. - no thorough evaluation of the object detection in cluttered scenes using the linear subspace method
Please provide the paper's structural or text errors (example: Figure 4 caption says z=x+y but Section II, paragraph 3 says z=x+a.).	page 1: "combining various descriptors[21,15], but a more" -> remove "but" page 2: LSM used without/before definition. "as following" -> "as follows" "is boosted up by" -> "is increased by" "on the other side" -> "on the other hand" "Alternative solution" -> "An alternative solution" page 3: "there is a redundancy that checks the same value" -> "there is redundant computation of the same value" page 4: "by following equation" -> "by the following equation" "in case of ColorCHLAC" -> "in the case of ColorCHLAC" "If we break up objects into reasonable sizes, its histogram" -> "If we break up an object into…" "normalize the final result" -> Clarify type of normalization (sum to 1?) Distinguish this from the later scaling of features to the range [0,1]. "but on the other side make the feature to the extent rotation-invariant and slightly slower" -> ?? Perhaps you mean "but on the other hand makes the feature rotation-variant and slightly slower"? "dice" -> "die" (in the singular case, like comparing cube to die) "descriptors which have additive property" -> "descriptors which have the additive property" "Owing to the property" -> "Owing to this property" "subspace of database object" -> "subspace of database objects" "use the top d dimensions" -> This is probably "t" from the following paragraph, but "t" occurs with different meaning in this paragraph. Introduce a consistent, unique variable for the number of top PCA dimensions used. (Caption for fig 3): You should identify this as _rotationally-invariant_ ColorCHLAC, not original ColorCHLAC. You should explain why the orange die matches the orange cube in the color histograms. Around here you say that "the additive property means that a global descriptor for an object cluster equals the summation of local descriptors of its sub-parts" but earlier you say "If we break up objects into reasonable sizes, its histogram can be approximated by the sum of the histograms of its sub-cells." Is this exact or an approximation? Why are objects being broken up into reasonable sizes? What is a reasonable size?

	"cubic subdivisions" -> Are these overlapping? page 5: SVM parameters paragraph is unhelpful without a bit more explanation. "shown in bottom part" -> "shown in the bottom part" "set of 49 artificially generated objects, 7 shapes" -> "…objects, consisting of 7 shapes…" "the fact that ColorCHLAC is not rotation invariant did not expose in the table" -> Objects are not rotationally symmetrical! Random rotation does affect results for ColorCHLAC! This needs to be explained much better. page 6: Caption for Table II should point out that this is on synthetic data. Also, the number of rows is too large. Also, a mention of why ColorCHLAC LSM beats ConVOSCH LSM in the first row would be helpful. "by avoiding putting of very different views" -> reword this

Masked Reviewer ID: Assigned_Reviewer_4

Review:

Question	
Please provide the technical review summary (write 1-3 full thorough paragraphs on technical contributions, strengths, and weaknesses.	This paper introduces two new descriptors that combine color and depth cues for object recognition. The features are histogram-based and extracted from a sliding-bounding box. A trained SVM or LSM classifier is used to determine whether the bounding-box encloses a known object. While the idea of combining color and depth information is not new (and I do not see a big distinction between "combining descriptors" and "developing ones that work in multiple features spaces"), the proposed feature extraction is fast and initial results on a 63 object dataset are promising. The authors plan to release their dataset, which is commendable. My main concern with the paper is that the experimental results are inconclusive. First, the differences between the various methods (features and classifiers) in the leave-one-out experiments (i.e., Table II) appear to be statistically insignificant. Second, the cluttered experiments are only run on a small number of scenes and no quantitative analysis given. In light of the scene shown in Figure 1, I find this very disappointing.
Please provide the paper's structural or text errors (example: Figure 4 caption says z=x+y but Section II, paragraph 3 says z=x+a.).	1. I found the introduction of the ConVOSCH and VOSCH descriptors in paragraph 4 ("In this paper...") very confusing. It would be more helpful to remove the parentheses and specify each descriptor separately (perhaps in a bulleted list with brief definition). 2. In section VI.B. I don't know what "did not expose in the table" means.
Additional comments to	The correct reference for "Integral Images" (section VI.D.) is Crow,

author(s)	Summed-area tables for texture mapping", SIGGRAPH 1984.