Joint Object-Material Category Segmentation from Audio-Visual Cues

Anurag Arnab, Michael Sapienza, Stuart Golodetz[1]
{anurag.arnab,michael.sapienza,stuart.golodetz}@eng.ox.ac.uk
Julien Valentin, Ondrej Miksik, Philip H. S. Torr[1]
{julien.valentin,ondrej.miksik,philip.torr}@eng.ox.ac.uk
Shahram Izadi[2]
shahrami@microsoft.com

[1] Department of Engineering Science
University of Oxford
Oxford, UK

[2] Microsoft Research
Redmond, US

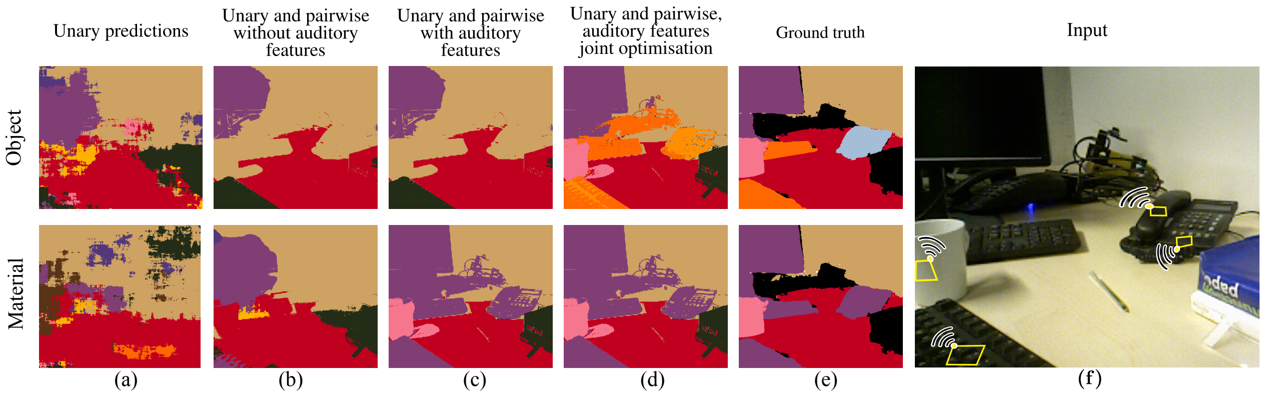
Unary predictions
Unary and pairwise without auditory features
Unary and pairwise with auditory features
Unary and pairwise, auditory features joint optimisation
Ground truth
Input
Object
Material
(a)
(b)
(c)
(d)
(e)
(f)

Figure 1: **(a)** The noisy predictions made by the per-pixel unary classifiers. **(b)** The output of the CRF using only visual features. **(c)** The use of auditory information improves material labeling. **(d)** Finally, joint optimisation between object and meterial categories improves object labelling as well. **(e)** The ground truth. **(f)** The input image, showing the locations where sound information is present.

It is not always possible to recognise objects and infer material properties for a scene from visual cues alone, since objects can look visually similar whilst being made of very different materials. In this paper, we therefore present an approach that augments the available dense visual cues with sparse auditory cues in order to estimate dense object and material labels. Since estimates of object class and material properties are mutually-informative, we optimise our multi-output labelling jointly using a random-field framework. We evaluate our system on a new dataset with paired visual and auditory data that we make publicly available. We demonstrate that this joint estimation of object and material labels significantly outperforms the estimation of either category in isolation.

By using sound, we are able to infer information about an object's material properties that would be difficult or impossible to obtain by visual means. This is evident from Figure 1 where we can see that the table, mug and wall all have similar local colour and texture, even though the table is made from wood, the wall from gypsum and the mug from ceramic. This leads to various object and material class labels being incorrect (Figure 1b). However, when we tap various objects in the scene and incorporate the resulting auditory information into our segmentation process, our predicted material label significantly improve (Figure 1c). We can then use these predictions to improve our object class predictions as well (Figure 1d).

Existing segmentation datasets do not provide audio-visual annotations as ground truth. Furthermore, it is not possible to simply augment them audio data, since we would need the original objects in the dataset to extract sound. As a result, we create our own dataset which we make publicly available[1]. In contrast to previous segmentation datasets, we annotate ours in 3D. We captured 9 different long sequences using a consumer-grade depth camera (ASUS Xtion Pro) and then reconstructed the 3D scene using the system of [4, 5]. This reconstructed scene was then annotated in 3D using an interactive scene segmentation framework [2]. This method allowed us to significantly decrease the annotation time since a typical sequence of 2000 frames could be fully annotated in about 45 minutes, which is far less than the 20-25 minutes per frame required to label each frame of the CamVid dataset by hand [1].

We captured our sound data using a portable condensor microphone (Samson GoMic). Due to the localised nature of sound, we can only associate sound data with the points at which the object was struck. This was done by annotating the approximate location at which the object was struck in the 3D reconstruction.

Since auditory information obtained by tapping objects is only available at sparse locations in an image, we need a method of propagating this information to the whole image. To this end, since estimates of object and material properties can be mutually informative, we use a two-layer CRF to model the joint estimation of object and material labels, and allow the two types of estimate to influence each other by connecting the two layers of the CRF with joint potentials. We minimise the energy,

$$E(\mathbf{x}|\mathbf{D}) = E^{O}(\mathbf{o}|\mathbf{I}) + E^{M}(\mathbf{m}|\mathbf{I},\mathbf{A}) + E^{J}(\mathbf{o},\mathbf{m}|\mathbf{I},\mathbf{A}), \qquad (1)$$

where $\mathbf{x}$ is an assignment to the random variable $\mathbf{X}$ that takes a label $[o,m]$ from the product label space of object and material labels, $\mathcal{O}\times\mathcal{M}$. The energy is conditioned on the visual and auditory data $\mathbf{D}=\{\mathbf{I},\mathbf{A}\}$. $E^{O}(\mathbf{o}|\mathbf{I})$ is the energy for the object labelling, conditioned on image data $\mathbf{I}$, $E^{M}(\mathbf{m}|\mathbf{I},\mathbf{A})$ is the energy for the material labelling, conditioned on image data $\mathbf{I}$ and audio data $\mathbf{A}$, and $E^{J}(\mathbf{o},\mathbf{m}|\mathbf{I},\mathbf{A})$ is the joint energy function linking the object and material domains.

The final joint energy function takes correlations between objects and materials into account, and encourages consistency between the two label categories. The joint potentials were learnt from the conditional distributions of the two labels in the training set. The first two energy functions consist of unary and pairwise potentials. The per-pixel unary potentials are obtained from a joint boosting classifier whilst the pairwise potentials takes the form of a mixture of Gaussian kernels to facilitate efficient mean-field inference [3, 6].

Using auditory data, the mean intersection-over-union (IoU) for material classification improves by 3.5% over the baseline which used only visual information. By employing joint optimisation between object and material classes, a further 4.1% improvement was obtained for object classification.

[1] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 2009.

[2] Stuart Golodetz, Michael Sapienza, Julien P C Valentin, Vibhav Vineet, Ming-Ming Cheng, Victor A Prisacariu, Olaf Kähler, Carl Yuheng Ren, Anurag Arnab, Stephen L Hicks, David W Murray, Shahram Izadi, and Philip H S Torr. SemanticPaint: Interactive Segmentation and Learning of 3D Worlds. Demo in SIGGRAPH ET, 2015.

[3] P. Krähenbühl and V. Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*, 2011.

[4] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (TOG)*, 2013.

[5] V. A. Prisacariu, O. Kähler, M. M. Cheng, C. Y. Ren, J. Valentin, P. H. S. Torr, I. D. Reid, and D. W. Murray. A Framework for the Volumetric Integration of Depth Images. *ArXiv e-prints*, 2014.

[6] Vibhav Vineet, Jonathan Warrell, and Philip HS Torr. Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. *IJCV*, 2014.

[1] http://www.robots.ox.ac.uk/~tvg/projects/AudioVisual/