2014 International Conference on Robotics and
Emerging Allied Technologies in Engineering (iCREATE)
Islamabad, Pakistan, April 22-24, 2014

Granting Space and Time to Words: Blind Semantic Spatial Localization for the case of Facial Images

Nikolaos Mavridis
Interactive Robots and Media Lab
NCSR Demokritos
Athens, Greece
nmav@alum.mit.edu

Thirimachos Bourlai
Department of Computer Science and Electrical Engineering
West Virginia State University
Morgantown, USA
thbourlai@mail.wvu.edu

***Abstract*— There exist a multitude of multimedia sources nowadays where audiovisual material is accompanied by labels. In many cases, the semantic content of these labels is not referring to the totality of the material they accompany; but only to a certain subset of it. Consider, for example, labels that refer to objects that are part of an image; or, alternatively, labels that refer to an event that is part of a video. Knowledge of the subset of the material to which the labels refer to can be very useful; for example, it can inform us regarding the detectability of the entity under the presence of selective occlusions or noise; or, it can be used to help segment out the referent from the material itself – and, among many other uses, to optimize the recognition of the entities that the words refer to. Towards these goals, in this paper we will present a method which allows such semantic spatiotemporal localization: given multiple instances of the material, and accompanying labels, we will produce subsets of the material which are most informative regarding the label; and which can be thought of as the spatiotemporally localized grounding of the concept represented by the words. The method is illustrated for the specific case of spatially localizing labels describing human faces or parts and artifacts of them; such as "beard", "glasses", "male", "old". No prior information about the spatial locus of the referents of these words is given; the algorithm blindly identifies the regions that are most informative for each label, and can be readily applied to robot vision.**

Keywords- face; semantics; localization; spatial

I. Introduction

With the ever-growing expansion of the internet as well as of electronic sensing devices (photo and video cameras, specialized imaging equipment etc.), a huge quantity of digital audiovisual material is being produced daily, in multiple forms, such as video clips, pictures, tomographic images, etc. Quite often, this material is being accompanied with words; labels describing it, and categorizing several aspects of it: For example, pictures uploaded to Picasa might have titles such as "Morning in Paris", or a short videoclip in Youtube might be accompanied by the text "Maradona's first goal". One of the possible questions that one could ask thus follows: "Is it the case that the text as a whole or its component words that accompanies an audiovisual material can be localized spatially or spatiotemporally within the material itself?" One possible version of this question becomes: "Which spatial or spatiotemporal region of the material is most informative about the label we are trying to localize?".

For example – for the case of "Maradona's goal" – which temporal fragment of the video clip, and which spatial locations within it contain the information required to be able to deduce that indeed "Maradona's goal" as an event partakes in it? Potentially, the temporal instants around the final kick and the crossing of the goal posts could be most informative; and the spatial locations around the ball, the feet of Maradona, his face or any region that enables us to identify him, as well as the goal post and line would be most informative. Of course, informativity requires a concrete mathematization; and one could think about many semantic variations to the meaning of the term: discriminative informativity (among a given universe of other labels), representative informativity etc.

There are many other cases, though, that a potential answer to the most informative spatiotemporal regions might not be that obvious as was the case in the "Maradona's goal" clip. Consider our first example – the photo labeled "Morning in Paris": which features of the photo would be most informative for the labels? If one is lucky, and a fragment of the Eiffel tower or the Pantheon is visible – as well as a morning sky or breakfast coffee – then the answer might be more straightforward. However, such highly discriminative features do not always exist; and the answer might not be so easy to find out through human intuition. Furthermore, and most importantly, what we are really after here is automatic spatiotemporal localization of meanings of words within audiovisual clips; and thus, although human intuition is useful in order to illustrate what we are after, it is not allowed to play any interventionist role within our method – which is required to be blind, i.e. fully automatic, without questioning or intervention of humans.

Now, imagine that indeed we have developed a method that satisfies the above requirements – an example of which we will illustrate in this paper. The important question that follows is: "But why could an answer to the question of which spatiotemporal regions are most informative for the label to be of any practical utility?". There are multiple potential benefits: for example, such a method can inform us regarding the detectability of the entities described under the presence of selective occlusions or noise; such information could be very useful towards many further goals – for example, towards customized selective cutting or compression of the material, while preserving recognizability. Furthermore, it can be used to help segment out the referent of the labels from the material itself, in order to be able to recompose them in different contexts – and, among many other uses, to optimize the recognition processes of the entities that the words refer to.

Thus, there are many beneficial uses to a method that can indeed spatiotemporally localize the most informative regions for given labels. There are also multiple computational challenges, though: If one was to consider, for example, all sets of possible subsets of the original material, then intractability would have immediately have arisen. Furthermore, one could envision that a generic method tackling many different types of labels and original material would not be easy to devise – and thus, more domain-specific solutions would be required.

In order to illustrate the method that we will present in this paper, we have selected a narrow yet clean and interesting domain: pictures of human faces, described by a set of labels – such as "old", "male", "asian", "has moustache" and so on. Assuming minimal prior structure in the labels (mutual exclusivity which leads us to postulate categories), we will proceed by examining the distributional structure and mutual informativity of the categories. Then we will reach our main goal: we will connect the labels with the facial pictures and specific pixels of them – and thus, discover specific loci of the face, which have high informativeness for the terms "old", "male", "moustache" etc. After presenting our results, we will discuss potential extensions of our method, also towards other types of labels and audiovisual material, before finally providing a succinct yet informative conclusion to our work.

II. Related Work

In the past, to our research, considerable research has taken place in a number of related areas: symbol grounding, grounding ontologies, spatial semantics, face recognition [1, 2]. However, nobody has directly tackled the question that we are asking here, that is blind semantic spatiotemporal localization, and more specifically apply it to the case of facial images and associated labels. Let us now consider the related areas where background work has been taking place in turn.

When trying to tackle the problem of lexical semantics by positing a meaning space which is non-linguistic, or when trying to connect symbolic representations that arise within an artificial agent to the world, the so-called "Symbol Grounding" problem [3] becomes central. Following the original statement of the problem, which has also appeared in different variations in the past, arguably going all the way back to ancient philosophy, there has been a stream of literature trying to investigate various sides of this problem. For example, some of the questions that have been investigated include: "How can create computational models of symbol grounding that enable robots or other artificial systems to understand the semantics of certain subsets of natural language, and to be able to utilize them while being embedded in the physical world or while being fed with sensory traces arising from the physical world, such as photos or videos?". Yet another question that has been investigated is: "How can groups of agents acquire grounded semantics, and how does meaning evolve in such a community?". Some classic examples of work related to the first question include [4], [5], where the subset of language that is tackled is related to spatial descriptions, including spatial prepositions such as "on top of", "inside", "to the left" etc. Other important projects related to spatial semantics include [6], where verbal directions for navigating a wheelchair are translated to trajectories, and [7], where a large corpus of descriptions of navigation directions was acquired, in order to build relevant computational models. Moving beyond spatial semantics, there exist projects such as [8], where a small robotic camera learns the grounded meaning of shape and color terms through examples [9, 10]. Here through the "grounded situation mode" architecture a manipulator robot is able to achieve capabilities comparable to those required to pass the "Token Test", a diagnostic test for detection of problems in the connection of words to sensory and motor abilities, which is administered to human children. Acquisition of computational models of grounded semantics through specially designed games is tackled in [11] for the case of massive online acquisition. Also, in [12, 13], the acquisition and processing of the world's online quasi-complete audiovisual corpus including a large percentage of the first three years of the life of a child are described, with a special focus towards building empirically-driven grounded models of meaning acquisition during child development.

Moving over to the second question, concerning models of the evolution of grounded meaning in groups of agents, the work of [14] and [15] is highly relevant. Also, other related work includes [16]. Extensions of the symbol grounding problem towards sets of symbols which partake in ontologies, are often described under the term "grounded ontologies". Beyond the connection of such nodes belonging to ontologies to their grounding domain, in this case, there also exist relations between such nodes with one another; for example, wordnet-like [17] meromorphic relations. A discussion of the relation of grounding to within-concept relational or similarity-based models is given in [18].

III. Methods and Results

In order to illustrate the problem and the proposed method, we have chosen the domain of pictures containing human faces, and descriptions of a set of labels. The question to be investigated is: how can we spatially localize the words contained in the labels, i.e. which regions of the picture are most informative towards labeling the picture with a specific word? Our picture set contained 3993 128 x 128 pixel 8-bit gray scale photos of faces, partitioned into a training set (Tr) of 1997 pictures, and a testing set (Te) of 1996 pictures. Each picture was accompanied with a verbal description containing a number of labels; in total, there were 19 words appearing as labels, namely L = {male, female, child, teen, adult, senior, white, black, asian, hispanic, other, serious, smiling, funny, moustache, beard, glasses, bandana, hat}. Not all images were of good quality; an example of two outliers that were part of the set is given in Fig. 1.

The first picture in Fig. 1, has been labeled "smiling white male child", while the second "serious adult white male moustache". While the labels are correct, the first picture has very extreme pose (head rotation), while the second contains a partial occlusion from the cigarette which hangs from the lips of the person depicted, and thus these pictures are outliers, which however often naturally arise within such datasets, and were thus not discarded in order to provide real-world results.

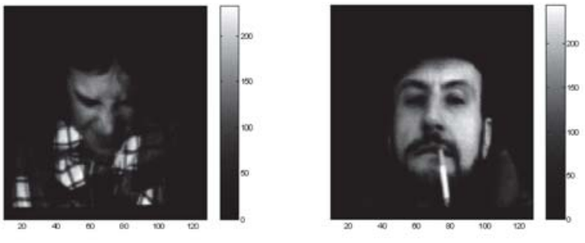

Figure 1. Two examples of outlier pictures from our dataset.

Regarding pre-processing, a face mask was derived by thresholding the average face, calculated as the sum image of all the faces, and an averaged resized image at 32 x 32 pixels was utilized for the spatial localization process, which will be described later, after exploring the probabilistic structure and mutual predictively of the verbal labels.

A. Investigating the Probabilistic Structure of the Labels

A label co-occurrence matrix was created, which confirmed that one can postulate Nd = 9 mutually exclusive dimensions of labels (Table 1).

TABLE I. DIMENSIONS OF MUTUALLY EXCLUSIVE LABELS

Dim	Description	Labels
1	Gender	L1 = {male,female}
2	Age	L2 = {child, teen, adult, senior}
3	Race	L3 = {white, black, asian, hispanic}
4	Expression	L4 = {serious, smiling, funny}
5	Moustache	L5 = {moustache, non-moustache}
6	Beard	L6 = {beard, non-beard}
7	Glasses	L7 = {glasses, non-glasses}
8	Bandana	L8 = {bandana, non-bandana}
9	Bandana	L9 = {hat, non-hat}

Thus, the number of categories corresponding to each label dimension were Ncat = {2,4,5,3,2,2,2,2,2}. Considering all allowable combinations, one can see that they belong to a discrete and finite 9D space. The distribution though of labels in this space is far from uniform: while the total number of possible combinations is 3840 = 2x4x5x3x2x2x2x2x2, there exist only 95 actual combinations, i.e. 2.474% of all possible. Of course, some of them are combinations which are naturally highly improbable; for example "female" and "beard". Others though, could be combinations which are possible, which however have not shown up in this particular dataset; for example, there were no "hispanic" and "children" in our dataset, which could well have existed in other datasets. Also, regarding the estimation of the probability of different label combinations, yet another consideration can be taken into account: that sometimes we expect pictures to be mislabeled; thus, when encountering a "female" and "beard" combination, this could also possibly be an indication of mislabeling. Finally, there is always the case of possible subjective inter-annotator disagreement regarding the meaning of labels; thus, a turquoise object might be thought of as being "green" by some, and "blue" by others.

Thus, if we know the hard constraints of the world in terms of allowable combinations of labels, and we are willing to hardcode them by hand, we can try to force the system not to take into account samples with incoherent label values. Else, we can let the system learn the allowable combinations based on what combinations it has encountered so far: thus, the system might attach a non-zero empirical prior to women with moustaches, which will eventually flatten out with a large enough training set. And, before having seen a hispanic child, it might assume that no such combination exists. However, this should not be treated as a "crisp" logical impossibility; but as a "soft" indication of low probability, that could well become non-zero on the basis of further observations of combinations – a situation which is similar to the treatment of "apax legomena" in traditional natural language processing.

So far, we have commented upon the co-occurrence structure of the labels, their empirical probabilities, and their embedding into a multidimensional space on the basis of their mutual exclusivity. Now, before proceeding to the connection of the labels to the pictures and the spatial localization of their meaning, let us delve a little deeper, starting from the following question: Given the 3840 potential combinations of labels, and the 95 actual combinations, how should one cluster together these combinations to categories, in order to assign them to classifiers? The number of possible partitions (sets of covering non-overlapping subsets) is huge. Another relevant question is: Q1) At the smallest possible granularity of partitioning, what are the 10 most likely of the 95 actual combinations of labels that were encountered? Yet another important question is: Q2) Before we proceed to connecting the labels with the pictures, how much mutual predictivity exists between labels? Let us start with the first question. The top ten cases of the training set, together with their priors are shown in Table 2.

TABLE II. TOP 10 LABEL COMBINATIONS, ACCOUNTING FOR 75%

Index	Probability	Combination
1	15%	female-adult-white-smiling
2	12%	male-adult-white-serious
3	11%	male-adult-white-smiling
4	10%	female-adult-white-serious
5	7%	male-adult-white-serious-withmoustache
6	4%	male-adult-white-smiling-withmoustache
7	4%	female-teen-white-smiling
8	4%	male-child-white-serious
9	4%	male-child-white-smiling
10	3%	male-teen-white-smiling

Notice that, by adding up the probabilities of the top 10 combinations, we have covered almost 75% of all cases; and the rest (25%) is covered by the remaining 85 cases. Also, notice how commonsense stereotypes have arisen: smiling white female adults, and serious white male adults dominate, for this specific dataset. Fig.2 shows the full priors of the training set label combinations in descending order.

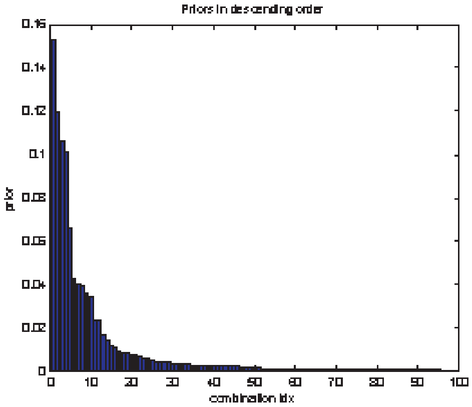

Figure 2. The Prior Probabilities of the Label Combinations of the Training Set, in Descending Order.

We visualize the marginalized priors across the 9 main dimensions of the labels, for the cases of the training and the testing set separately (Fig. 3 and Fig. 4).

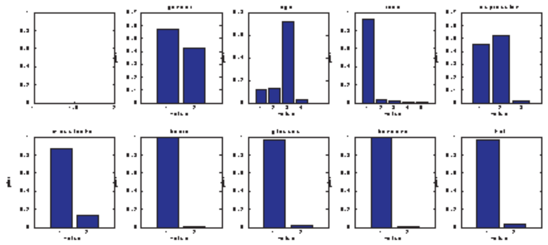

Figure 3. The Marginalized Priors of the Training Set.

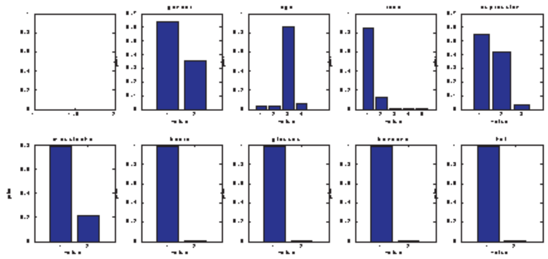

Figure 4. The Marginalized Priors of the Testing Set.

By observing Fig. 3 and Fig. 4, the following points come up:

- Asymmetry: Every label except gender/expression has very imbalanced priors, much more for the last four (high entropy)

- There is considerable difference of the marginalized priors between the train and test sets, and so they are somewhat mismatched for recognition, even at a very fundamental level. Taking into account the size of the samples (2K), this difference seems significant statistically. Other evidence might also signal this; for example, mean face and eigenvector differences, or mutual recognition rates etc.

- By the trivial classification rule of just choosing the value having the largest prior of the training set, we can get on the testing set the following recognition accuracies across the 9 dimensions: {0.6398, 0.8667, 0.8512, 0.5496, 0.7876, 0.9950, 0.9960, 0.9960, 0.9900}. These are quite high percentages, and thus any results should be considered above this baseline.

B. *Mutual Predictivity between Labels*

Now, let us proceed to the second question that we had set out, namely Q2: How much mutual predictivity exists between labels? I.e. if we know that one label is “male”, what can we say regarding the probability of other labels?

Let us start by considering pairs of labels, and their marginalized priors, so we can gain some insight in combining outputs of pairs of classifiers (instead of going to the full 9D problem). For example, just by detecting zeros in the 2D priors, we can easily through some lines of code automatically generate statements of the form (with all the caveats given above regarding zero empirical priors):

Facts I know about the relation of_gender_with_beard:
-> Noone_with_female_gender_also_has_a_beard!-

Such statements might be worth exploiting, if for example we have a highly robust moustache classifier correcting a mediocre gender decision and vice versa, or even in more balanced cases given not highly overlapping errors. This argument can go a bit further (considering pairs again), and the mutual information of their priors shows such “inbalanced” cases that can easily be exploited (again depending on their individual classifier performance and matching). For pairs, we get the result in Fig. 5 where the upper figure denotes color coded log of mutual, for pairs of labels. The lower figure denotes the thresh holding of the above.

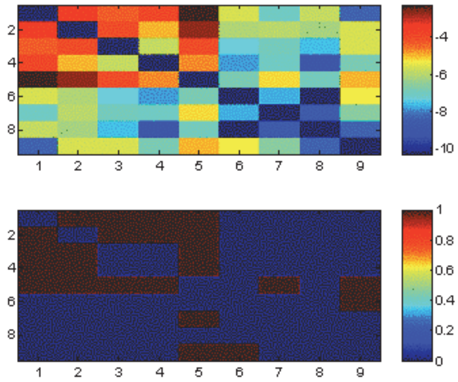

Figure 5. Mutual Information of Priors.

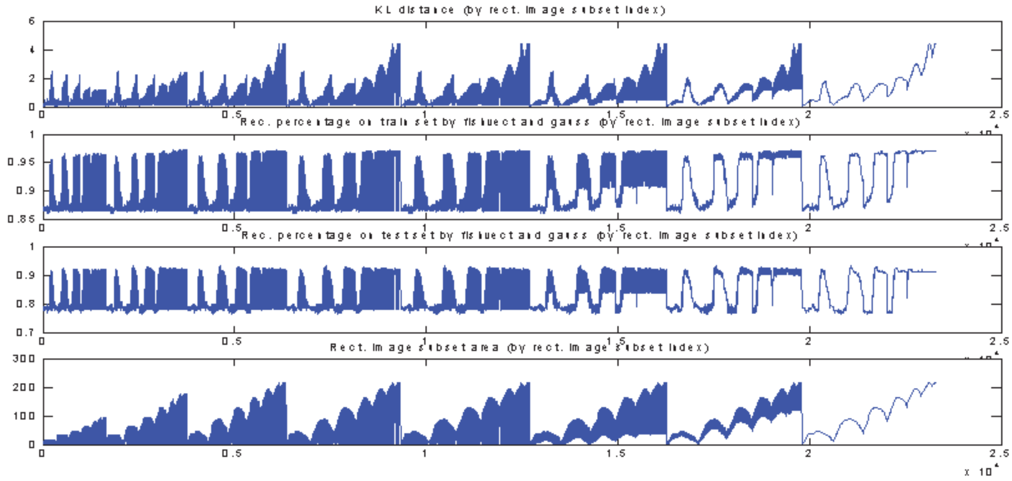

Figure 6. KL, % on Train, % on Test, Area in Pixels, Arranged by Subset Index (0-23K), for the Case of Label 6 (Moustache).

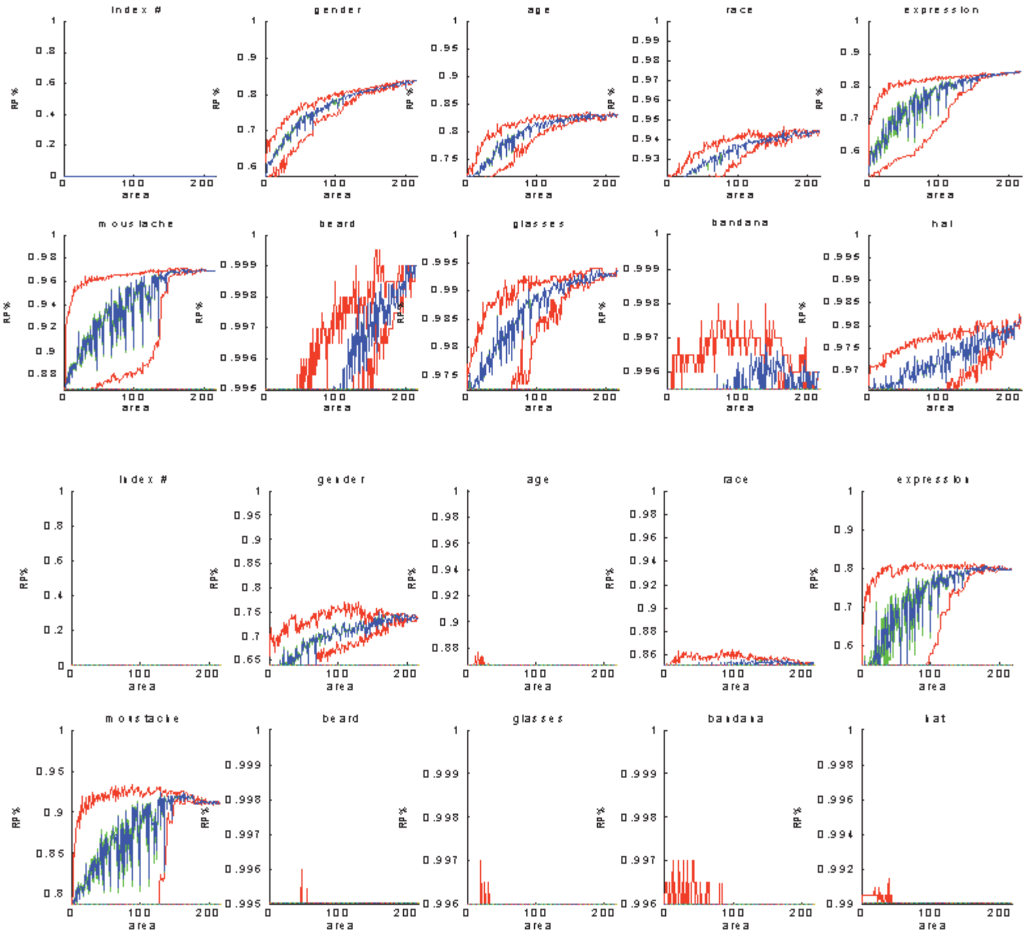

Figure 7. Recognition Percentage of Simple Classifier on the Training Set (Above) and on the Testing Set (Below).

It is thus apparent which pairs of features offer useful information about each other: (1,5)=(gender, moustache) has the highest score, but also (expression, gender) etc. Now, let us move over to our main question: how can we spatially localize the meaning of the labels?

C. Semantic Spatial Localization of the Labels

In simple words, here we are attempting to answer the following question: for a specific label (for example, moustache), which pixels (most probably forming areas) of the image contain the crucial information for recognition? Or, else, can this thing labelled "moustache" be localised somewhere in the face, based on the training data? As discussed in the introduction, there are multiple benefits of obtaining an answer to this question.

One of the first issues that needs to be addressed, when attempting an answer, has to do with the set of possible regions that can serve as a potential answer. What are we looking for? Any possible subset of pixels of the image? The number of possible subsets is huge; for the 32 x 32 image, we have 2 to the power of 1024 subsets; and one cannot expect to be able to exhaustively search in all of them. Thus, we chose a subset of all possible subsets: namely, those that are rectangular, with sides belonging to {4,8,12,16,20,24,28,32}. Later, in the second stage of the process that we will describe, we can also derive softer regions with more complicated shapes; but initially we work with rectangular regions.

Then, following the set of subsets choice, one needs to choose a criterion for selecting among the subsets, the one that contains enough crucial information, while discarding the larger subsets. We have chosen two such criteria: the first is Kullback-Leibler distance (KL distance), and the second is the recognition percentage of a fast-and-simple classifier based on fisher discriminants.

For the first criterion, a symmetrized version (arithmetic mean) of the KL distance was used, divided by the number of pixels in the subset, to counterbalance for large numbers of pixels. Also, the resistor-average KL [19] was tried out, but for the gaussian case we used, the later it is equivalent. The implementation of the KL distance was problematic: early histogram-based approaches suffered from zeros, and while some quick fixes for the 1D case were successful, but these didn't generalize well to the multi-D case. Thus, to get a quick working fix, we estimated means and covariance matrices and used the KL formula for gaussians [19]. However, there were severe numerical problems in this case too, due to the often ill-conditioned inversion of covariances. For the second criterion, our simple classifier consisted of a 1D fisher vector projection (discriminating label with maximum prior with the union of the rest), followed by gaussian fitting (based on empirical sigma and mu for each class), empirical priors, and the associated quadratic decision boundary.

Now let us move on to results. There exist 23K rectangular regions with sides belonging to {4,8,12,16,20,24,28,32} in the 32 x 32 pixel image. If we arrange them linearly, and calculate: KL divergence, recognition percentage on the training set (i.e. without generalization), recognition percentage on testing set (i.e. true percentage with generalization), and subset area in pixels, we get diagrams in Fig. 6. By observing Fig. 6, one can see that:

- Large-scale and small-scale periodicities due to the looping of rectangle side sizes

- There is certainly a relation between KL and recognition percentage, of course even more so on the train set. However, this relation is not always monotonic; there exist pairs of increasing KL values with decreasing percentages. This was to be expected however, as KL is related to asymptotic performance, furthermore under other idealizations.

In general, recognition performance of the simple classifier on the train set was found to be a more reliable predictor of the performance of more complicated classifiers on the test set than KL, although further experimentation and quantitative justification would be required for a stronger statement regarding this observation. Even if the original subset areas should be multiples of 16 (4 x 4), because the regions of the subsets were furthermore masked by the almost-elliptical face mask, all possible areas between one and the total mask size emerged, with a spiky decreasing histogram of areas. If we now arrange the subsets by area, we can find the statistics of the criteria as a function of subset size (Fig. 7) where statistics is considered as a function of feature set size (red = min and max, blue = mean, green = one sigma boundaries), y-axis denotes ABOVE MAX. PRIORS. Points to notice in Fig. 7: First, we do get satisfactory separation in almost all labels – moustache seems best, and race worst. Second, there are labels for which adequate information can be found even in small areas of the picture (i.e. are highly localised); moustache is such a prime example, and also expression. It gets to 90% of its best percentage with a 32-pixel subset, and probably even smaller (if we allowed all possible rectangle side sizes we would know). i.e.:

```
'---> LABEL=moustache'
Best KL values [0.50821,4.4972], firstmin@=1, firstmax@=216, but we still
get over the acceptable=4.0475, with an area as small as =211
Best TE perc [0.78758,0.93387], firstmin@=1, firstmax@=89, but we still get
over the acceptable=0.91924, with an area as small as =16
```

- There are other labels where we continue to get more useful information as the subset size increases, and we should probably use the whole picture if we can, such as gender. i.e.:

```
'---> LABEL=gender'
Best KL values [0.12138,0.96421], firstmin@=2, firstmax@=9, but we still
get over the acceptable=0.86779, with an area as small as =9
Best TE perc [0.65731,0.77054], firstmin@=1, firstmax@=130, but we still
get over the acceptable=0.75746, withan area as small as =90
```

- Of course, highly localized labels have more variance for a fixed size k for small sizes.

In order to provide a clearer picture of what is happening across the 9 dimensions of labels, we can create a visualization of the best feature sets, as well as those at 90% of maximum performance but with minimum size (Fig. 8). Notice that indeed the highly localized features require small sets, and also notice how the inherent symmetry of the human face functions: in many cases, only half the area is required. We certainly get what we expected for moustache, expression, glasses, bandana and hat.

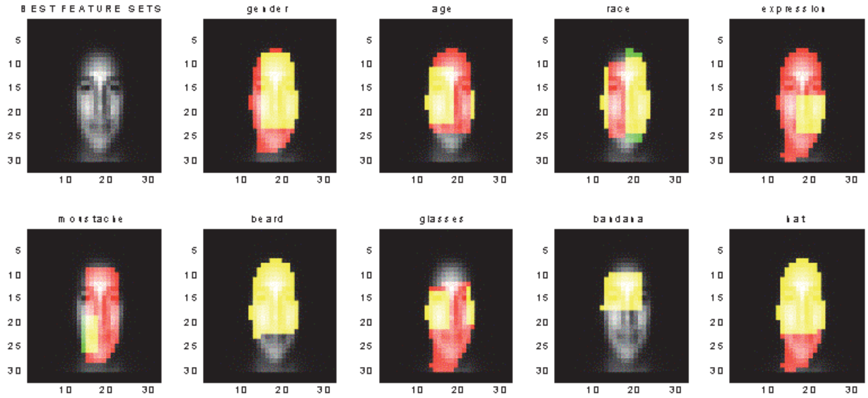

Figure 8. Best Feature Sets and Satisfactory Sets (90% of Best Performance).

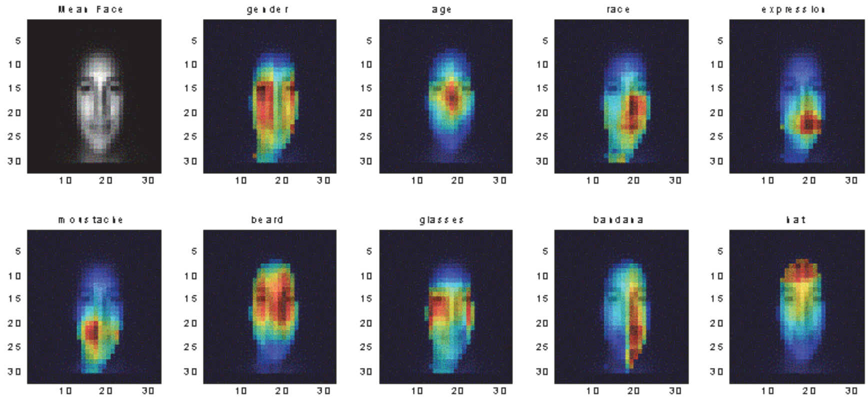

Figure 9. Relative Frequency of Pixel Belonging to Best Performing Sets for Size k.

Now, in order to be able to view not only regions, but also a distribution of informational importance of different regions, as it pertains to their frequency of inclusion in highly informative subsets for a specific label, we can try the following visualization. We plot the number of times a pixel belongs to a best performing (for size k) feature set over the number of times it belongs to all sets created (Fig. 9). Here, times pixel belonged to a best performing subset for an area k. k=1:maxk over times pixel belonged to a feature subset that was evaluated, where "best performing" is according to recognition percentage on train set by the simple classifier. There are several points worth noting in Fig. 9.

Semantic localization is now evident. We have indeed found out which part of the image is more informative towards the labels EXPRESSION, MOUSTACHE, GLASSES and HAT. For results with inadequate rec. performance: the BANDANA result is spurious, due to the fact that two of the supposedly bandanas are mufflers worn on the neck, and the beard result is very general due to the suspicious non-labeling of beards (when a moustache existed,).

Thus, through the process that we have described, we have successfully localized the most important areas for each label, and thus can justify selecting these. And we have also roughly grounded the meaning of words as "moustache" in spatial regions of the pictures. In summary, we have taken the following steps: we started by investigating the probabilistic co-occurrence structure of the labels, and embedding them into a nine-dimensional space. Then, after discovering hard constraints and visualizing mutual predictivity of the labels, we moved on to our main question: how are the labels connected to spatial regions of the pictures? We chose a set of possible subsets for region (rectangular regions with specific side lengths), and two metrics for informativeness of the subset towards recognizing the category of the label (KL-divergence as well as simple classifier performance), and after these choices were made we asked: how is the informational importance of different regions distributed, as it pertains to their frequency of inclusion in highly informative subsets for a specific label? To answer this question, we plotted the number of times a pixel belongs to a best performing (for size k) feature set over the number of times it belongs to all sets created, in

Fig. 8: which provides a clear visualization of the automatic spatial semantic localization of the labels, thus achieving the goal that we had set out for.

IV. Discussion of Future Steps

There exist multiple avenues for extension of our system and concept. First, we plan to apply our method for other domains, beyond labels of faces, also to audiovisual material where we will also be able to provide not only spatial but also temporal localization. Second, we plan to investigate alternative choices for set of subsets to be investigated, as well as generative mechanisms which adjust sets of subsets on the fly; for example, by appropriate adaptation of sequential forward selection and other such methods [20].

Finally, we plan to chain the resulting output of our method to task-specific compression and recognition subsystems, in order to investigate the potential gains to be made by adaptively providing particular importance to the most informative task-related spatiotemporal regions, which our method produces given a set of labeled material.

V. Conclusion

Large quantities of labeled audiovisual material that exist today, either in the internet or in other sources. Knowledge of the subset of the material to which the labels refer to can be very useful; for example, it can inform us regarding the detectability of the entity under the presence of selective occlusions or noise; or, it can be used to help segment out the referent from the material itself, and among many other uses, to optimize recognition of the entities that the words refer to.

Motivated by the above state of affairs, in this paper we presented a method which allows such semantic spatiotemporal localization: given multiple instances of the material, and accompanying labels, our method produces subsets of the material which are most informative regarding the label; and which can be thought of as the spatiotemporally localized grounding of the concept represented by the words. A detailed illustration of the internals and the details of our method was given through the specific case of spatially localizing labels describing human faces or parts and artifacts of them; such as "beard", "glasses", "male", "old" and so on. No prior information regarding the spatial locus of the referents of these words is needed for our method; the algorithm blindly identifies the regions that are most informative for each label.

While presenting our example, multiple relevant questions were asked, and computational answers were provided – for example, regarding the probabilistic structure of the labels and their mutual informativity. Our method was able to finally deliver a clear visualization of the most informative loci regarding our labels, and thus to give a clear indication of its power. Finally, as discussed, multiple avenues for extension exist; and the basic principles of semantic spatiotemporal localization can easily extend beyond the case of facial regions which we have illustrated in detail in this paper, and can be thus successfully applied to many other interesting domains.

Acknowledgment

We would like to thank Rosalind Picard and Niloy Mukherjee for their valuable assistance in this project.

References

[1] G. Toderici, S. M. O'Malley, G. Passalis, T. Theoharis and I. A. Kakadiaris, "Ethnicity- and gender-based subject retrieval using 3-D face-recognition techniques," *International Journal of Computer Vision*, vol. 89, pp. 382-391, 2010.

[2] Z. Zeng, T. Fang, S. K. Shah and I. A. Kakadiaris, "Personalized 3D-aided 2D facial landmark localization," in *Proceedings of the 10th Asian conference on Computer vision - Volume Part II*, 2010, pp. 633-646.

[3] S. Harnad, "The symbol grounding problem," *Physica D*, vol. 42, pp. 335-346, 1990.

[4] T. Regier and L. Carlson, "Grounding spatial language in perception: An empirical and computational investigation," *Journal of Experimental Psychology: General*, vol. 130, pp. 273-298, 2001.

[5] K. R. Coventry and S. Garrod, *Saying, Seeing and Acting: The Psychological Semantics of Spatial Prepositions*. Psychology Press, 2004.

[6] S. Tellex, *Grounding Language in Spatial Routines*. Master Thesis, Massachusetts Institute of Technology, 2006.

[7] A. Anderson, M. Bader, E. Bard, E. Boyle, G. M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson and R. Weinert, "The HCRC map task corpus," *Language and Speech*, pp. 34351-34366, 1991.

[8] D. Roy and A. Pentland, "Learning words from sights and sounds: A computational model," *Cognitive Science*, vol. 26, no. 1, pp. 113-146, 2002.

[9] N. Mavridis and D. Roy, "Grounded situation models for robots: Where words and percepts meet," in *IEEE International Conference on Intelligent Robots and Systems*, 2006, pp. 4690-4697.

[10] N. Mavridis, *Grounded Situation Models for Situated Conversational Assistants*, Phd Thesis, Massachusetts Institute of Technology, 2007.

[11] P. Gorniak, *The Affordance Based Concept*, PhD Thesis, Massachusetts Institute of Technology, 2005.

[12] J. Orkin & D. Roy, "The restaurant game: Learning social behavior and language from thousands of players online," *Journal of Game Development*, vol. 3, no. 1, pp. 39-60, 2007.

[13] D. Roy, R. Patel, P. D. Camp, R. Kubat, M. Fleischman, B. Roy, N. Mavridis, S. Tellex, A. Salata, J. Guinness, M. Levit, P. Gorniak., "The human speechome project," in *Proceedings of the 28th Annual Cognitive Science Conference*, 2006.

[14] Luc Steels, "Semiotic dynamics for embodied agents," *IEEE Intelligent Systems*, pp. 32-38, 2006

[15] L. Steels and F. Kaplan, "Bootstrapping grounded word semantics," in *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, Cambridge University Press, 2002.

[16] J. M. Siskind "Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic", *Journal of Artificial Intelligence Research*, vol. 15, pp. 31-90, 2001.

[17] Christiane Fellbaum, *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[18] J. J. Bryson, "Embodiment versus memetics," *Mind & Society*,vol. 7, pp. 77–94, 2008.

[19] D. Johnson and S. Sinanovic, "Symmetrizing the Kullback-Leibler Distance," *IEEE Transactions on Information Theory*, 2001.

[20] F. J. Ferri, J. Novovicova, J. Kittler, "Floating search methods for feature selection with nonmonotonic criterion functions," in *12th IAPR International. Conference on Pattern Recognition*, 1994, pp.279-283.