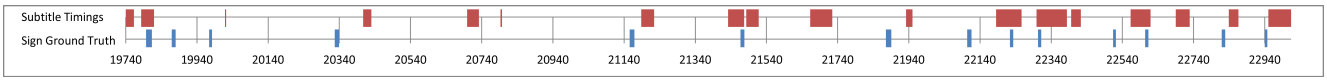
Learning Signs from Subtitles: A Weakly Supervised Approach to Sign Language Recognition

Helen Cooper and Richard Bowden
CVSSP, University Of Surrey.
Guildford, UK
{H.M.Cooper, R.Bowden}@surrey.ac.uk

(a) Alignment over a section of video. Of the 23 words in the subtitles only 7 are present in the sign gloss.



(b) Alignment for the word/sign Army/Soldier. For the 18 subtitles there are only 14 sign occurrences. In two cases the sign is >200 frames away.

Figure 1. Examples of correlation between subtitles and signs.

Abstract

This paper introduces a fully-automated, unsupervised method to recognise sign from subtitles. It does this by using data mining to align correspondences in sections of videos. Based on head and hand tracking, a novel temporally constrained adaptation of apriori mining is used to extract similar regions of video, with the aid of a proposed contextual negative selection method. These regions are refined in the temporal domain to isolate the occurrences of similar signs in each example. The system is shown to automatically identify and segment signs from standard news broadcasts containing a variety of topics.

1. Introduction

This paper proposes a fully-automated, unsupervised method to learn sign by correlating broadcast video with the associated subtitles. The paper presents a novel temporally constrained adaptation to data mining which employs efficient pruning strategies to find similarities in sections of video. Accuracy is further increased by using contextual negatives in the mining process and results are shown in both the task of word spotting and the more complex task of sign-subtitle alignment via iterative temporal refinement.

One of the biggest challenges to face anyone approaching the problem of Sign Language Recognition (SLR) is the lack of labelled realistic data. Sign Language, being as complex as any spoken language, has many thousands of signs each differing from the next by minor changes in hand motion, shape or position. The grammar of sign also modifies signs according to what is being said. While the handshape for the sign 'Aeroplane' remains constant, the motion will change depending on the context *e.g.* 'taking off' or 'landing'. This, coupled with the intra-signer differences, make true SLR an intricate challenge. The majority of available data sets contain non-co-articulated sign or short sentences, recently this has been supplemented with the Boston NC-SLGR data set [3] containing 15 short stories. Nevertheless, video sign corpus is scarce and creating new data sets is non-trivial since the subjects should be native signers and hand labelling has to be completed by someone competent in the language. There is, however, a vast amount of data being broadcast daily with an inset signer. This offers a potentially limitless source of data, signed by native signers and covering a wide range of topics; unfortunately, it lacks any ground truth labelling. What it does contain are subtitles which show some correlation to what is being signed as shown in figure 1. This figure shows both a sign gloss and the corresponding transmitted subtitles. Examples of the footage are shown in Fig 3. As can be seen, the word order and grammar differ between English and Sign, which results in the correlation being weak and there is rarely full alignment. This is to be expected when one considers that the subtitles paraphrase the audio content while the sign is a translation of that content. The region around a subtitle may contain a sign, or several instances or none at all. Though

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

the correlation is strong enough that there are similarities in signs over several subtitle examples it presents a difficult challenge to any learning algorithm.

1.1. Related Work

It has been shown that tracking a signer's head and hand positions can allow a sign to be described [12][13][11] even when the training examples are limited [8]. However, these approaches require ground truth labelled data and with the exception of Kadir *et al*., have a very limited vocabulary in part due to the lack of data available.

Farhadi and Forsyth approached the idea of alignment between sign and English subtitles [6]. They use HMMs with both static and dynamic features to get estimates of the start and end of a sign before building a discriminative word model to perform word spotting on 31 different words over an 80000 frame children's film. Their data appears to have a one-to-one mapping and matching order between the signs and the subtitles which they use to help them remove false positives. While some sign languages may exhibit such mappings this is not true of sign languages in general as previously shown in Figure 1.

In the computer vision community, there has been a move toward larger datasets and weak supervision for learning to allow the use of freely available information such as flickr photos or Google image searches. Recently the concept of data mining has been introduced to the vision community grouping together SIFT features for object recognition [10], to cluster together near duplicate images in large data sets [5] and to combine low level corner features for action recognition in videos [7]. Proving itself in these situations to be a strong tool for picking discriminate features, it lends itself to the idea of finding the similarities in multiple sections of video whilst discarding the noise and irrelevant data.

2. Methodology

Our approach consists of several different stages as shown in figure 2. In the first instance, a tracking system is used to obtain head and hand positions, these co-ordinates are then clustered to obtain a quantized description for each frame which are then temporally concatenated to create bi-frame features. Using these features, the similar sections of each video block can be mined and the responses used to show where a sign is likely to be located. When using this method for word spotting, the parameters required for mining are experimentally chosen without difficulty. However, when the problem is extended to weakly supervised learning from subtitles, the parameter selections becomes more difficult due to the lack of prior knowledge about the data. The proposed solution runs the mining at a series of parameters, combines the results using Mean Shift to give new potential sign positions before iteratively refining in the temporal domain to give more definitive results. The mining uses both positive examples found from the subtitles and some contextually chosen negatives to aid in removing frequently-appearing non-target signs. More details on each of these stages are given in the following sections.

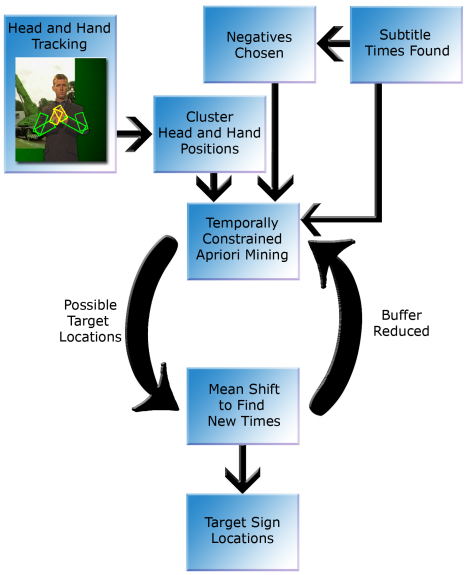


Figure 2. System overview

2.1. Tracking and Quantization

In order that signs may be recognized, it is necessary to have a descriptive input to the mining algorithm. To this end, the tracking system described by Buehler *et al*. [4] is used which employs a generative model based approach to tracking the motion of the upper torso.

Using pixel based head and hand positions as features for mining would result in a large feature set with no generalization. By quantizing the positions using K-means clustering, the feature set is reduced to 10 possible head positions and 20 positions for each of the hands. The number of clusters was chosen by examining the cost graph for different cluster sizes to determine an optimal number. This provides 3 features per frame each able to take 10 or 20 values. This is expanded to include temporal information by concatenating the features on each frame with those of the previous, forming simple visual bigrams similar to those used in speech recognition. In order that mining can distinguish between the head and hands, each symbol is prefaced by a feature type. A symbol S_n is always a 5 digit feature made up of the single digit feature type F_t, the 2 digit feature F of that type for the frame n and for the frame $n-1$. For example, if in frame n, the head position is assigned to cluster 4, the left hand to cluster 11 and the right hand to cluster 8, and in frame $n-1$ the head position is assigned

to cluster 4, the left hand to cluster 10 and the right hand to cluster 12 then the symbols assigned to frame n would be: 10404 21110 30812.

$$\begin{aligned} F_t &\in \{1,2,3\} \\ F_n, F_{n-1} &\in \{1\dots10\} or \{1\dots20\} \\ S_n &= (10000 * F_t) + (100 * F_n) + F_{n-1} \end{aligned} \quad (1)$$

2.2. Mining

Apriori mining in its original form [1] returns an exhaustive list of all commonly occurring rules in a set of given examples. Originally used in market data research, it's often referred to as market basket analysis, extracting correlations from people's shopping baskets to enable the supermarkets to identify related products. Rules take the form of a set of symbols (the antecedent) implying another symbol (the consequent), *e.g.* $A \Leftarrow B, C, D$. Apriori mining can be easily manipulated to exclude rules which also occur in a set of negative examples. This is done by including a positive or negative identifying symbol in each example, then setting the mining to return only rules with the positive identifier as the antecedent e.g. $Pos_{id} \Leftarrow 10404, 21110, 30812$. With large sections of video, an exhaustive list of rules includes many which combine temporally distant symbols. It is therefore prudent to use some form of temporal information during mining to remove rules which cannot describe a sign as they contain elements which are too far apart. One obvious answer is to use Sequential Pattern Mining [2] however, natural signers rarely repeat signs in an identical manner *e.g.* the head may be tilted just before or during, the left hand may rise alongside the right or it may happen a second after. This is especially noticeable when co-articulations are taken into account. This leads to the conclusion that temporal bagging would be a viable alternative, enforcing a temporal coherence between features without rigidly stating a specific order.

2.3. Temporally-Constrained Apriori Mining

Mining traditionally takes a combination of parameters which dictate the strength of the rules returned, the support $S_{A \Leftarrow B}$ and the confidence, $C_{A \Leftarrow B}$.

$$\begin{aligned} S_{A \Leftarrow B} &= P(A, B) \\ C_{A \Leftarrow B} &= P(A|B) = \frac{P(A, B)}{P(B)} \end{aligned} \quad (2)$$

For Temporally-Constrained Apriori Mining, 3 parameters are required: the minimum number of positive examples which must display the rule mP which is directly related to the support; the maximum number of negative examples which are allowed to exhibit the rule mN, related to the confidence; and the temporal distance allowed between symbols mT. As with traditional Apriori mining, the first step is to reduce the symbol set to only those that, on their own, can meet the minimum positive criteria. Trees are then built of all possible rules containing these symbols, pruning out branches which don't meet the positive support requirement since, despite which symbols are included, a child node can never have a greater support than its parent. The trees built can be very large, due to the number of possible symbols and the frequency with which they occur, therefore a tractable implementation is required.

Rule trees can be calculated and assessed recursively so the maximum memory in use at any one time is governed by the depth of the tree and not the width. There are 2 types of pruning that can be accomplished easily on the fly. The first terminates any branch that does not meet the minimum positive criteria as mentioned previously. The second, terminates branches to a rule containing a symbol which has a value less than the greatest value in the current rule. While this produces unbalanced trees, it stops duplicates from occurring. *e.g.* a rule $Pos_{id} \Leftarrow 21009, 21110$ would not branch to rule $Pos_{id} \Leftarrow 21009, 21110, 10404$ since $10404 < 21110$ and 21110 is the largest value in the parent, but it could branch to $Pos_{id} \Leftarrow 21009, 21110, 30812$ since $30812 > 21110$. At each branch, a rule which meets both the minimum positive and maximum negative conditions is written to a file and can therefore be deleted. By building multiple rule trees each starting with a different symbol matching the positive requirement criteria, the algorithm lends itself to multi-threading for use on multi-core CPUs since each tree can be built by a different thread.

As in the original Apriori Mining algorithm, each rule has a confidence. When evaluating a block of video, a sliding window of size mT is applied across the example and the confidences of all rules appearing in the sliding window are summed to give a response for that window. The peak response can then be found and that section of video is labelled as the region containing most similarities to the positive examples used in mining.

2.4. Using Contextual Negatives

Given the noisy input data that will be used, it is imperative that negatives are chosen carefully, however due to the scale of the problem, they also need to be found automatically. Ideally they should contain similar content to the 'noise' in the positive examples, then mining can find what symbol sets belong to the target sign alone. Mensink and Verbeek use a similar idea when searching for images of people [9]. They look for people who appear frequently alongside the target subject and exclude them from the query search. In the case of subtitles, the process is similar, words which appear in the same section of subtitles as the target word are accumulated. A subtitle search is then performed for these words to create a negative data set. This negative set should be contextually similar to the

Figure 3. Three examples of the sign army/soldier, the top 2 rows contain unmodified signs the bottom row is a modified version where the head and hand positions differ from the unmodified version.

noise in the positive data set i.e. it will contain similar signs to the non-target signs in the positive set. The final step is to exclude any examples in the negative set which are temporally too close to those in the positive data set, these are most likely to contain the target sign and so should not form part of the negative set.

2.5. Localising Signs - Mean Shift

The parameters chosen for mining will severely alter the rule set found, if mP is set too high it will result in no rules being chosen, if mP is set too low or mN is set too high, the rules found will be meaningless. When word spotting, this isn't an issue since the sign is known to occur once within an example and negatives are known not to be contaminated, so parameters can be found experimentally. However, when trying to learn signs from subtitles, the problem is less well-defined. A block of video around a subtitle may contain the sign, it may contain multiple instances or it may contain none. Furthermore, the negatives cannot be guaranteed to be uncontaminated with instances of the target sign. Subtitles containing the desired word may also be close enough to each other that when a buffer is applied to either side, they may overlap. In addition to this, each word/sign combination will exhibit a different correlation pattern so a generic rule cannot be applied to calculate the desired parameters. Examples of subtitle/sign correlation are shown in figure 1.

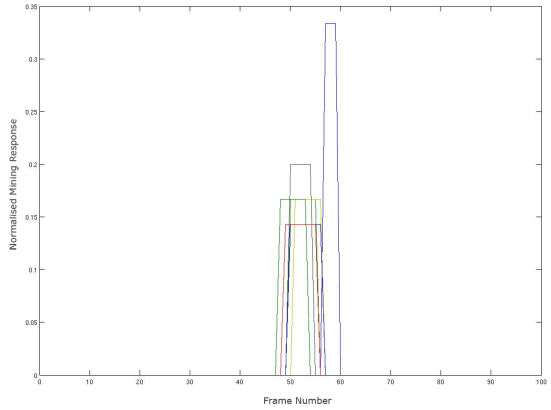
A solution is to run the Temporally-Constrained Apriori Mining with various parameters and then draw conclusions about the sign positions from the combinations of the responses. After each mining stage is performed, a peak response is found in each positive example and for each set of parameters. A histogram of the frames in which these peaks occur is then constructed. Since the examples used sometimes have overlaps or are temporally close to each other, it is necessary to combine the responses across the examples in their original temporal situations. The top line of figure 5(a) shows the histogram built for the sign 'Army/Soldier'. It can be seen that there are several small groupings of peak responses that should each return a single new starting point. To combine these groups, a Mean Shift algorithm is applied to find the modes of the data. Kernel centres are initialized on each of the non-zero bins and are shifted by calculating moments of bins within the kernel to the centre. In this case a moment is defined as the bin's weight multiplied by its distance from the kernel centre. The kernel centres are required to sit on a non-empty bin and kernels which overlap by more than 4/5 of the kernel size are combined into a single kernel.

3. Experimental Results

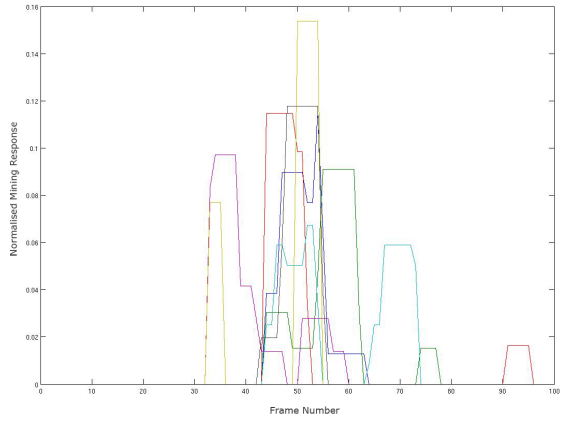
In order to assess the ability of mining to select relevant feature combinations, the first task was to locate signs in a more constrained manner. To this end, word spotting experiments were performed. After the concept had been proven on the constrained data, alignment between sign and subtitles can be approached.

3.1. Word Spotting

Two signs were chosen to be ground truthed in a half hour news program, the sign for 'Plane', relating to an Indonesian plane crash, and the sign 'army/soldier' (the sign is the same for both words). These signs were chosen because there were several occurrences, most of which were signs un-modified by context, although modified signs were included in the data (*e.g.* plane crash vs plane). Three examples of the signs for army/soldier are shown in figure 3. There were 10 examples of army/soldier and 7 examples of plane, both tests used the same set of 11 negatives chosen manually from temporally distant areas of the data to ensure no contamination with the target signs. These signs are each around 10 frames long. A buffer of 50 frames was applied either side of the ground truth label to give sections of video $\simeq$110 frames long. The 'Army/Soldier' data was mined with $mT = 10$, to correspond to the known sign length. Different values for mP and mN were assessed ranging from 0% to 100% in 10% intervals. The cleanest result was gained using $mP = 70\%$ and $mN = 0\%$. This is unsurprising since the negative data was chosen specifically not to contain any examples of the sign so mN should be low and the positive examples were all know to con-



(a) 'Army/Soldier'



(b) 'Plane'

Figure 4. Mining responses across examples of sign Army/Soldier and Plane when $mP = 70\%$ and $mN = 0\%$. Each line shows a different example which contains the ground truth and a buffer of 50 frames applied to either side. Therfore a peak around frame 50 constitutes a correct identification and a peak elsewhere is an incorrect identification.

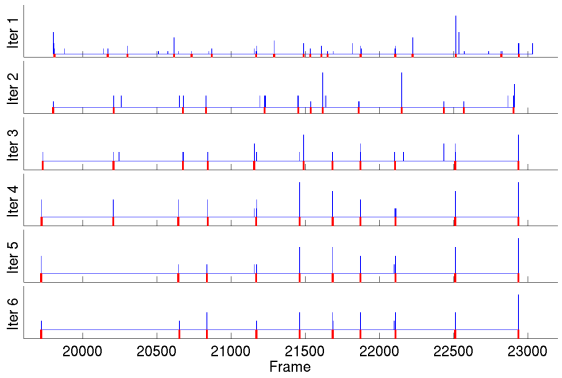
tain examples of the sign (though not all identical) so mP should be relatively high. A reduced number of test were performed on the 'Plane' data with $mP = 50\% \ldots 100\%$ and $mN = 0\% \ldots 30\%$. The same values of mP and mN showed the best results. The response graphs for these 2 tests are shown in figure 4. In both cases the peaks occur around frame 50 where the signs are know to start. Whilst not every example of the sign is found (scoring 6/7 for plane and 7/10 for army/soldier) the false positives are low with only one being found in plane where the peak occurs 14 frames before the sign.

3.2. Weakly Supervised Learning from Subtitles

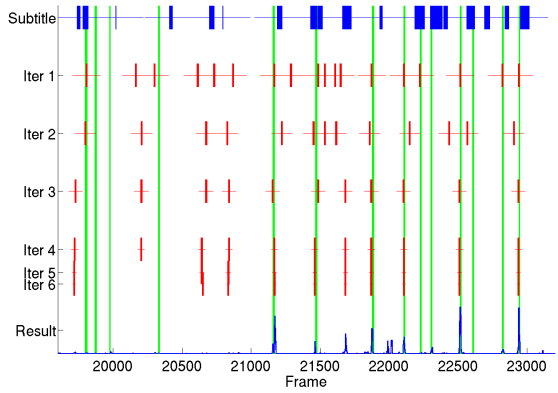
Since sign language and English have different grammars and structures, the first challenge was to find a list of words that could be aligned. Most frequently occurring words in the English language (*e.g.* articles, pronouns, prepositions and conjunctions) do not have a rigid equivalent in sign. All the tests were run on the same 30 minute (46207 frame) news broadcast and target words were selected that were mainly nouns and which appeared in the subtitles a minimum of 4 times. Some English words were grouped together since they are expressed by the same sign (*e.g.*'Army' and 'Soldier' or 'Obese' and 'Overweight'). Tests were run using negatives randomly selected from the remaining video or using the contextual negative selection method. The iterative temporal honing applied a buffer of 200 frames either side of the subtitle to begin. Followed by 100, 75, 50, 25, 15 and 10 frames around the selected regions for each of the successive iterations. Mean Shift used a kernel size of 200 at each iteration. The parameters for mining were $mP = \{40\% \ldots 100\%\}$ and $mN = \{0\% \ldots 10\%\}$ in 10% steps with a check that $mP > 2\ examples$. Examples of the iterative process results are shown in figure 5. Figure 5(a) shows the kernel centers Mean Shift picks for each of the modes of data and figure 5(b) shows how these modes alter with each iteration, it can be seen that the number of modes reduces as the iterations close in on the sign. After the iterative temporal refinement, the final mining responses of each parameter set were summed on a frame by frame basis to give a view of the full video. The results of which are shown in Figure 6. With 1000 random negatives the response for 'Soldier' shown in figure 6(a) peaks on 3/14 possible signs and has 8 false positive peaks. When using 590 contextually chosen negatives as in figure 6(b) it peaks on 7/14 possible signs and has only 1 false positive peak. If this is expanded to include sections of subtitle labelled army as well as soldier figure 6(c) then it peaks on 9/14 possible signs and has 3 false positive peaks. Also shown are the responses for the words 'Weight'(figure 6(e)) and 'Obese'(figure 6(d)) independently and then when combined with 'Overweight' (figure 6(f)) which is only present twice in the subtitles. It can be seen that 'Obese' performs poorly on its own but when combined with the stronger 'Weight', the results are less noisy than either of the original words, this is in part due to the increased number of examples available but also due to reducing the occurrences of signs, which are similar in form to the target sign, appearing in the negative examples set. Also shown in figures 6(g) to 6(i) are the responses when the classifiers are run over the entire video, note how there are very few false positive peaks firing outside the region containing the signs. The top five signs were all over

# Words	23	20	15	10	5
Mean	53.7%	58.4%	68.0%	79.7%	91.1%
SD	26.1%	24.7%	20.6%	13.8%	6.0%

Table 1. Correct positive responses within the original subtitle buffer.



(a) Histograms produced after each iteration with the Mean Shift kernel centers shown in red below the axis.



(b) Subtitles (top row in blue), kernel centers (following rows in Red), buffers (horizontal lines) for each iteration and the ground truth (green vertical lines). The summation of the mining responses from the final iteration is shown along the bottom in blue.

Figure 5. Iterative temporal refinement for the sign 'Army/Soldier'

90% correct see Table 1. While this drops as the number of examples mined is increased, the accuracy only drops below 70% when 15 signs are considered. Overall 23 words were tested (chosen since they occured 4 or more times in the 30 minute video), and the mining was able to isolate signs on average 53.7% of the times they fired within the original subtitle buffers.

The time taken to learn a rule set for sign detection varies depending on the complexity of the problem. Given the pre-tracked data the entire process from extraction based on subtitles, through the iterative process and to the final recognition stage typically takes about an hour per word on a machine with 4 dual core 3GHz P4 processors.

4. Conclusions

Having introduced an adaptation to Apriori mining, which makes it suitable for mining sections of video, it has been shown that Temporally Constrained Apriori Mining is a good method for locating and segmenting signs in large sections of video. This has then been extended to work with a weakly-labelled, noisy data set. The addition of a contextual negative data set to increase performance has demonstrated the concept of using subtitled, inset-signer broadcasts to automatically identify, classify and segment sign without ground truth data.

5. Future Work

It has been found that the biggest problem when moving from word spotting to subtitle sign alignment is the noise increase in the data set. Ideally, data should contain many repetitions of the same words in various different contexts, however, this rarely happens in real broadcasts. Instead, the data is often contextually blocked, sun and rain usually appear together in the weather report for example. Increasing the number of broadcasts used would give a substantial boost in the number of words which could be mined. Current endeavours are to increase the video corpus significantly. Including appearance based features would lead to further improvements, since sign characteristics such as hand shape and facial expression could be coded to allow better discriminative models to be built.

6. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 231135 - DictaSign and grant agreement no 215078 - DIPLECS.

References

[1] R. Agrawal and T. Imielinski. Mining association rules between sets of items in large databases. In *ACM SIGMOD Conf. on Management of Data*, pages 207–216, 1993.

[2] R. Agrawal and R. Srikant. Mining sequential patterns. In *Research Report RJ 9910*, pages 3–14, 1995.

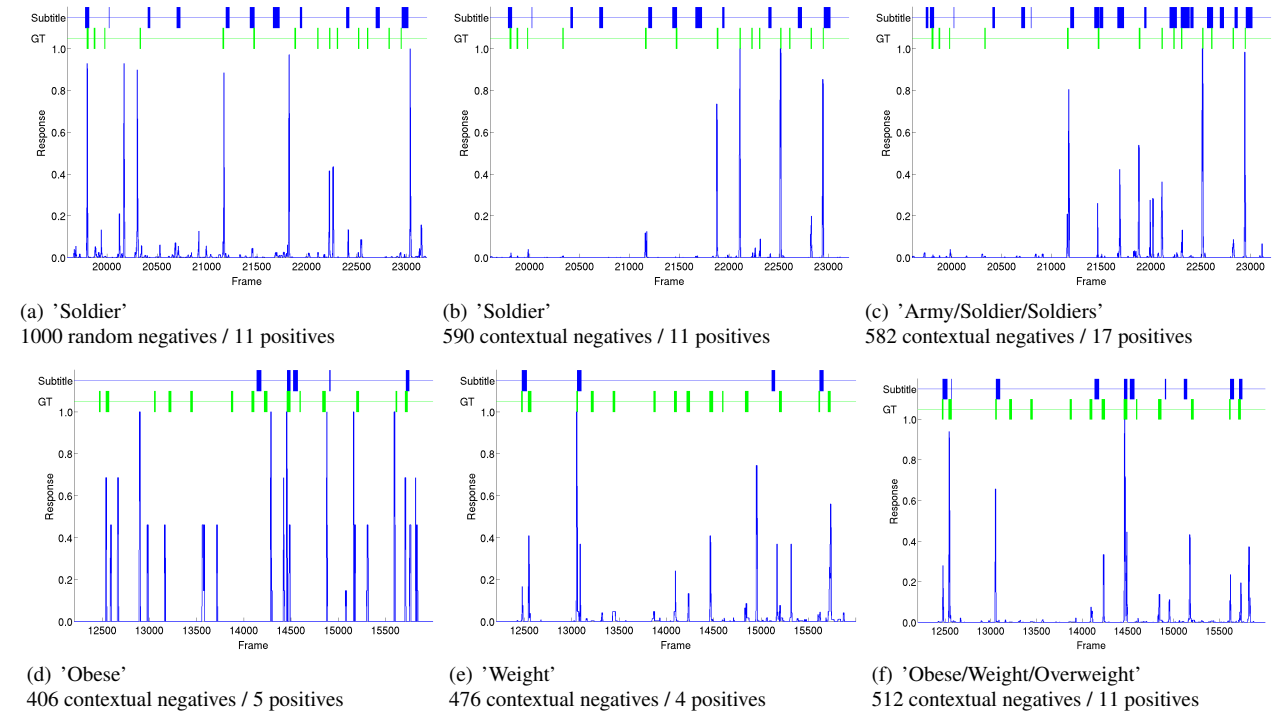
[3] Boston University. National Centre for Sign Language and Gesture Resources - American Sign Language Linguistic Research Project. http://www.bu.edu/asllrp/cslgr/, 2006.

[4] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. Long term arm and hand tracking for continuous sign language TV broadcasts. In *Proc. of the British Machine Vision Conf.*, 2008.

[5] O. Chum and J. Matas. Web scale image clustering, large scale discovery of spatially related images. Technical report, CMP, CTU, Prague, 2008.

[6] A. Farhadi and D. Forsyth. Aligning ASL for statistical translation using a discriminative word model. In *CVPR '06: Proc. of the 2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 1471–1476, Washington, DC, USA, 2006. IEEE Computer Society.

[7] A. Gilbert, J. Illingworth, and R. Bowden. Scale invariant action recognition using compound features mined from dense spatio-temporal corners. In *ECCV (1)*, pages 222–233, 2008.



(a) 'Soldier'
1000 random negatives / 11 positives

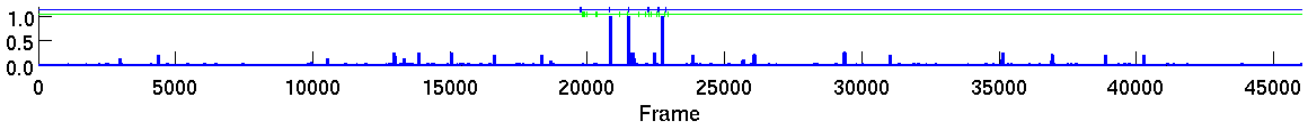
(b) 'Soldier'
590 contextual negatives / 11 positives

(c) 'Army/Soldier/Soldiers'
582 contextual negatives / 17 positives

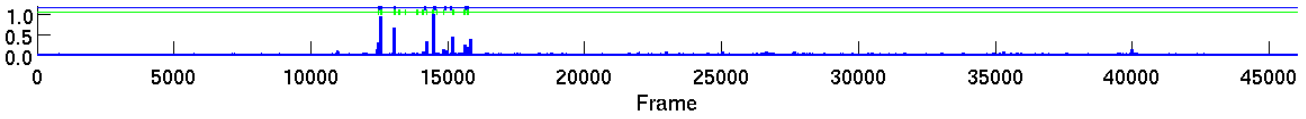
(d) 'Obese'
406 contextual negatives / 5 positives

(e) 'Weight'
476 contextual negatives / 4 positives

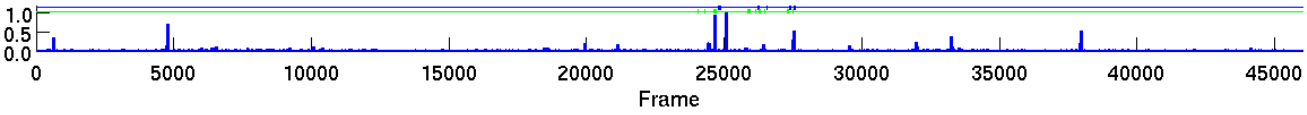
(f) 'Obese/Weight/Overweight'
512 contextual negatives / 11 positives



(g) Responses across the 46207 frame video for the sign 'army'



(h) Responses across the 46207 frame video for the sign 'Obese/Weight/Overweight'



(i) Responses across the 46207 frame video for the sign 'school'

Figure 6. Responses of the rules found by mining after refining iteratively in the temporal domain, the original subtitle times are shown across the top in blue, the ground truth below them in green.

[8] T. Kadir, R. Bowden, E. J. Ong, and A. Zisserman. Minimal training, large lexicon, unconstrained sign language recognition. In *Proc. of the British Machine Vision Conf.*, volume 2, pages 939–948, Kingston UK, Sept. 2004.

[9] T. Mensink and J. Verbeek. Improving people search using query expansions: How friends help to find people. In *European Conf. on Computer Vision*, 2008.

[10] T. Quack, V. Ferrari, B. Leibe, and L. Van Gool. Efficient mining of frequent and distinctive feature configurations. In *Computer Vision, 2007. ICCV 2007. IEEE 11th Int. Conf. on*, pages 1–8, 2007.

[11] A. Shamaie and A. Sutherland. A dynamic model for real-time tracking of hands in bimanual movements. In *5th Int. Gesture Workshop*, pages 172–179, Genova, Italy, Apr. 2003.

[12] T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. In *ISCV '95: Proc. of the Int. Symposium on Computer Vision*, page 265, Washington, DC, USA, 1995.

[13] C. Vogler and D. Metaxas. ASL recognition based on a coupling between HMMs and 3D motion analysis. In *Proc. of the Int. Conf. on Computer Vision*, pages 363–369, Mumbai, India, Jan. 1998.