How to Frame a Video: Contextual Thumbnail Selection in Dutch

Anonymous ACL submission

Abstract

Since its introduction, CLIP has been used in many different image-text matching tasks. This paper aims to apply the benefits of CLIP to the thumbnail selection task, utilizing CLIP's ability to find matching image-text pairs. Finding a suitable thumbnail for a video can be a tedious and time-consuming task, especially when the thumbnail should be representative of the video's content. In this paper, we introduce a pipeline that leverages CLIP and contextual text from the video to select a suitable frame as a thumbnail. Experiments show that OpenCLIP implementations achieve superior performance on Dutch data compared to other CLIP and vision-language models. Furthermore, using the title and the video synopsis as textual input outperforms other pipelines that use summaries, only the title, or a combination of the similarity score with face and text detection scores. Compared to different reference systems, our approach also shows superior results according to human evaluators.

1 Introduction

With the number of videos uploaded to different platforms on the internet every hour, it is important to quickly and effectively capture users' attention to stay competitive and relevant in the market. On YouTube alone, over 20 million videos are uploaded every day (YouTube). The first point of contact for a user with a video is often the thumbnail, as it is the first thing they see. Therefore, a thumbnail that accurately reflects the video's content is important for catching the user's attention and keeping the user engaged. For online news and entertainment videos, having thumbnails that accurately represent the video's content is important to help users understand what the video is about and keep them engaged on the site. Furthermore, manually selecting a thumbnail for the number of videos produced each day for media companies is a time-consuming and tedious task. Thumbnail selection can be automated using face detection, text detection, or image aesthetic and technical scores to find frames that contain faces, have no text, and are considered nice images (e.g., no blur). But these techniques do not take into account what the video is about and may therefore not be very representative of the actual content of the video. Therefore, thumbnail selection tools that incorporate video content in their selection process play an important role. These factors highlight the need for an automated thumbnail selection pipeline that considers not only aesthetic features but also contextual ones, producing context-aware thumbnails without human intervention.

To the best of our knowledge, the only comparable work using CLIP for thumbnail selection is (Fantini, 2024), where the author uses extracted keywords for semantic similarity as part of a larger pipeline, rather than experimenting with different text inputs. In comparison, this work focuses explicitly on eliminating the need for additional data, such as audio or aesthetic scores, aiming to use only CLIP's capabilities to find pleasing images. Furthermore, all previous work focuses on English data, while our research focuses on Dutch data.

The main contributions of this paper are:

1. The utilization of CLIP models for the thumbnail selection process, eliminating the need for a huge amount of labeled data.
2. Fine-Tuning a CLIP model on a specific domain to potentially improve the results, still resulting in a way smaller dataset needed.
3. Applying CLIP to non-English text to select thumbnails for videos in Dutch.
4. Creating different frame selection methods that fit the text spotting bias of CLIP.
5. Human expert evaluation comparing a CLIP-based pipeline against a proprietary pipeline

and reference systems, showing the capabilities of CLIP to select a thumbnail on Dutch videos.

All of the code is publicly available on GitHub [1].

2 Related Work

Previous work on automatic thumbnail selection can be divided into two categories: including or excluding textual context to select a fitting thumbnail.

Different approaches that do not include textual context exist in the literature. (Song et al., 2016) cluster the frames after filtering them to determine the relevancy for every frame according to its cluster size, and then selecting the frame with the highest aesthetic score for each cluster ranked by cluster size. Other work, such as (Husa et al., 2022), focuses on aspects like close-up shots, face detection, and logo detection, as well as blur detection and image quality prediction to rank the candidate frames. In comparison, (Apostolidis et al., 2023) uses reinforcement learning to train a network on a manually labeled dataset to select the best thumbnail based on aesthetic and visual features.

Early work on incorporating textual input was done by (Gao et al., 2009). Here, the authors use keywords to extract images fitting the theme of the video from Google Image Search, then rank the frames according to the color theme of the retrieved images. One of the first approaches to leverage neural networks for the task of thumbnail selection was proposed by (Liu et al., 2015). There, the authors train visual-semantic embeddings using a CNN and the "GloVe" model to select the thumbnail closest to a user's query. Later work introduces a model that first filters the frames according to their aesthetics and then extracts features from the frames, audio, title, and description of the video, creating a combined vector which is then compared to the filtered frames to select the one with the smallest mean squared error (Yu and Shi, 2021). (Fantini, 2024) uses CLIP to match frames to extracted keywords, building a pipeline that not only includes text-image matching but also focuses on faces and emotions, and crops frames.

CLIP (Radford et al., 2021) is a vision–language pretraining approach that learns aligned image and text representations. The goal is to decrease the distance between matching text-image pairs while increasing the distance between non-matching pairs. CLIP produces a shared feature space that yields promising zero-shot results across various downstream tasks. In our work, we utilize the joint representations of CLIP to find a fitting thumbnail for a video title or synopsis.

3 Methodology

This section outlines the methodology for multilingual evaluation, as well as the CLIP pipeline setups and automatic as well as human evaluations.

3.1 Multilingual Model Evaluation

Different implementations of CLIP (Radford et al., 2021), OpenCLIP (Ilharco et al., 2021), ALIGN (Jia et al., 2021), and LiT (Zhai et al., 2022) already provide out-of-the-box multilingual support. However, the capabilities of the models on Dutch data need to be evaluated, as performance can differ greatly across languages. Table 1 gives an overview of the evaluated models.

All models were evaluated on the Flickr30k-200 and XTD200 datasets from CLIP Benchmark (LAION-AI, 2024) in English and Dutch, as these datasets already provide prompts in both languages. The goal of both benchmarking datasets is to retrieve the image with the highest similarity to a given query.

As a metric, the average image retrieval recall at 1, 5, 10, and 20 is used across both datasets for each language.

3.2 CLIP Pipelines

As one of the OpenCLIP models achieves the best results on Dutch data, see Section 5.1, different pipelines, including different OCR methods, text inputs, and quality frame filtering, are set up using this model. Table 2 shows the various experiments performed. The following sections provide a more detailed explanation of the implemented approaches.

3.2.1 OCR

Two different text-detection variations are implemented. One that detects text across the whole image, except the upper right corner, because that is where a logo is burned in and would always be detected as text, and another that crops the image to the bottom middle to only detect whether the frame has subtitles. Both methods work well to

[1] https://anonymous.4open.science/r/How-to-Frame-a-Video-Contextual-Thumbnail-Selection-in-Dutch-559E/README.md

Model	Source	Reason	Recall@5
ViT-L-14-336, openai	CLIP OpenAI	best performing CLIP model according to the original CLIP paper	0.41
ViT-L-14, commonpool_xl_s13b_b90k	OpenCLIP	as comparison to the CLIP model	0.83
nllb-clip-large-siglip, mrl	*OpenCLIP*	*best multilingual according to OpenCLIP evaluation*	*0.91*
xlm-roberta-base-ViT-B-32, laion5b_s13b_b90k	OpenCLIP	standard multilingual model in OpenCLIP	0.85
nllb-clip-base-siglip, mrl	OpenCLIP	2nd best multilingual according to OpenCLIP	0.77
xlm-roberta-large-ViT-H-14, frozen_laion5b_s13b_b90k	***OpenCLIP***	***best performing multilingual ViT according to open clip***	***0.92***
ViT-B-16-SigLIP-i18n-256, webli	*OpenCLIP*	*after xlm the best performing ViT according to OpenCLIP*	*0.86*
ViT-H-14-378-quickgelu, dfn5b	OpenCLIP	best open clip model for classification	0.92
ViT-SO400M-14-SigLIP-384, webli	OpenCLIP	best open clip model for retrieval	0.79
M-CLIP/XLM-Roberta-Large-Vit-B-16Plus	Multilingual-CLIP	best performing m-clip model	0.91
M-CLIP/XLM-Roberta-Large-Vit-L-14	Multilingual-CLIP	2nd best m-clip model	0.87
kakaobrain/align-base	ALIGN		0.61
LiT-L16L	LiT	the one that is used on their GitHub page	0.20
LiT-B16B_2	LiT	the one that is used on their GitHub page, which has better retrieval results	0.18
LiT-B16B	LiT	the standard LiT model	0.17

Table 1: Evaluated Models on English and Dutch. **Model** describes the model checkpoint used. **Source** specifies the library/technology the checkpoint is taken from. **Reason** gives more explanation about why this checkpoint was included in the evaluation. **Recall@5** shows the mean over both Dutch datasets. **Bold** denotes the best performing model on Dutch data. *Italic* shows the models with a Recall@1 ≥ 0.7.

eliminate frames containing text in the given image area. The main reason a subtitle-only version is implemented is that logos or other letters in the image that are not proper text are detected as text, leading to many frames being discarded and often leaving only a few or none. As the graphics with names do not pose a big problem for the aesthetics, it was decided that removing only the subtitles is sufficient. Further technical considerations can be found in Appendix D.

The reason the frame filtering is important is that CLIP is particularly drawn to images with text. Different papers (Lin et al., 2024) (Materzyńska et al., 2022) show that in CLIP, image encoders can match an image of a text with natural images of the same scene. Therefore, without proper prefiltering of the images, the top images would almost always contain text, which should be avoided. The frame filtering method, which only removes subtitles, proved useful for improving image input to the models, finding a balance between discarding frames with unwanted text without removing too many frames to leave none.

3.2.2 Translation

Because the performance on English data is slightly better than on Dutch data, as can be seen in Figure 2, further evaluation is done by translating the titles into English to use as textual input. To evaluate this approach, the captions of the two benchmarking datasets are translated from Dutch to English using the MADLAD-400-3B-MT model (Kudugunta et al., 2023), which is already available on the infrastructure. The reason for not translating the captions beforehand is to simulate the real-life pipeline, where the data is also available only in Dutch and would need to be translated first.

3.2.3 Summary

The video titles used as model inputs can be relatively long and linguistically verbose. To mitigate potential noise introduced by redundant lexical content, we create summaries of the titles and synopses using GPT-4o. The prompt constrains the generated summaries to a length of 50 characters.

Figure 1 shows the attention maps for the full-length and summarized video titles. The visualization of the attention indicates that a substantial fraction of tokens in the full-length title receives very low attention weights, suggesting limited direct interaction with other tokens in the self-attention mechanism. In contrast, the attention map for the summarized title shows fewer low-attention tokens, while retaining most tokens that receive high atten-

ID	System	Filtered	OCR	Context	Extra Info
A	**Baseline**	**-**	**-**	**Title**	**-**
B	Baseline	-	-	Title	+1.15× similarity score
C	**Baseline**	**-**	**E-SUB**	**Title**	**+1.15× similarity score**
D	CLIP	-	-	Title	-
E	CLIP	-	-	Translation	-
F	CLIP	-	-	Title + Synopsis	-
G	CLIP	-	P-TXT	Title	-
H	CLIP	-	E-TXT	Title	-
I	CLIP	-	E-TXT	Title	-
J	CLIP	-	E-SUB	English Title	-
K	CLIP	-	E-SUB	Title	-
L	**CLIP**	**-**	**E-SUB**	**Title + Synopsis**	**-**
M	**CLIP**	**-**	**E-SUB**	**Summary (50 characters)**	**-**
N	CLIP	Quality	-	Title	-
O	CLIP	Quality	P-TXT	Title	-
P	CLIP	Quality	E-SUB	Title	-
Q	CLIP	Quality	E-TXT	Title	-
R	**CLIP**	**Quality**	**E-FB***	**Title**	**-**
S	CLIP	-	E-SUB	Title + Synopsis	Fine-Tuning

Table 2: Overview of evaluated baselines and CLIP variants. **System** indicates if the current system is used as described in Appendix B, or the CLIP variant is used. For **Filtered** '-' all extracted frames were used, for 'Quality' only frames above the technical and aesthetic score threshold are used, see 3.2.4. **OCR** describes if OCR is used as described in Section 3.2.1, where '-' means no OCR. 'P-TXT' indicates PyTesseract removes frames with text. 'E-TXT' indicates 'EasyOCR' removes frames with text. 'E-SUB' indicates EasyOCR removes frames with subtitles. E-FB is a special case that indicates EasyOCR text filtering with a fallback to EasyOCR subtitle filtering (ID K) if no frames pass the initial filtering. **Context** defines the text input given to the CLIP model. **Bold** rows indicate the models used in the CLIP Pipeline evaluation.

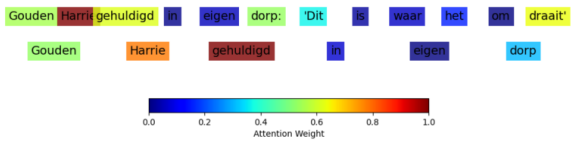

Figure 1: Top: Text Attention on the Full Video Title. Bottom: Text Attention on the Summarized Video Title

tion in the original input.

These observations suggest that the summarization process preserves the semantically salient concepts while reducing lexical redundancy. Consequently, title shortening appears to improve input efficiency without discarding information that is most relevant to the model's attention mechanism.

3.2.4 Quality

The current pipeline gives a technical and aesthetic score for every frame. The higher the score, the higher the frame is rated in the current pipeline. To incorporate these scores into CLIP experiments, it is decided to filter the frames beforehand and remove those with low technical or aesthetic scores. To find a threshold for what constitutes a low score, the distributions of the technical and aesthetic scores for the Top 5 and non-Top 5 frames in the current system for the test videos are plotted. The distributions can be seen in Appendix E.

In the end, the selected thresholds for the technical and aesthetic scores are 4.7 and 5.7, respectively. When filtering the frames according to these criteria, only those with a technical score greater than 4.7 and an aesthetic score greater than 5.7 are considered.

3.2.5 Fine-Tuning

The best model according to human evaluation 3.3.1 is fine-tuned on domain-specific data. Therefore, over 34,000 header images and title pairs from news and entertainment articles are used. As these header images are selected by humans, they give a good indication of what describes an article well.

The model is fine-tuned using LoRA. Only the "query", "dense", and "out_proj" weights are fine-tuned with a LORA rank of 2 and an alpha of 4. Furthermore, Lora dropout is set to 0.1, and the learning rate is 10^{-5}. The model is trained for 50 epochs, and the best checkpoint, the one with the lowest validation loss, is taken as the final finetuned model.

3.3 Evaluation

Because there is no ground truth, meaning no frame that is *the best* thumbnail for a video, as this is a highly subjective topic that would always be biased towards the person creating a dataset, the compari-

son between the systems is made by human evaluators. First, non-expert human evaluators evaluate the results of 4 different CLIP pipelines against the current baseline. Afterwards, the best CLIP model is evaluated by experts against other thumbnail selection implementations.

3.3.1 CLIP Pipeline Testing

To compare the different systems, it is necessary to let human evaluators pick the best thumbnail from a selection. Therefore, an experiment is set up that presents the evaluators with the top pick of the following systems:

- Baseline (Current System) (ID A)
- Similarity Score + E-SUB + Title (ID C)
- E-SUB + Title + Synopsis (ID L)
- E-SUB + Summary (ID M)
- Quality + E-FB + Title (ID R)

The above five systems are selected because they use different approaches to filter frames beforehand and show promising results in the Top 5 results. The prompt given to the evaluators can be seen in Appendix F. To prevent a bias from the position a system is at for every experiment, the order of the frames is shuffled. This is shown in Appendix F. Furthermore, if two or more systems pick the same frame and the evaluator selects one of them as the best, all other systems that picked the same frame are also counted as picked.

3.3.2 Expert Testing

The best performing CLIP pipeline according to human evaluation described in Section 3.3.1, a fine-tuned version of it, the current system, the implementation from (Lin et al., 2024), and the implementation from (Vasudevan et al., 2017) are presented in a similar set-up as before but this time to a group of experts consisting of 5 professionals working on media content. These experts are asked to pick the thumbnail they like most and indicate whether they think it would be an acceptable thumbnail for the video.

4 Data

This section describes the different kinds of textual input and frame filtering.

4.1 Image Data

The current proprietary system already has an algorithm that samples the video frames at equally spaced time intervals, since using every frame would not be feasible. Additionally, many frames that appear one after another in the video look very similar. Therefore, sampling them in fixed intervals reduces the amount of duplicate/similar frames. In addition to the frames, the current pipeline also supplies technical and aesthetic scores, text detection, face detection, and face clustering for each frame.

4.2 Text Data

For each video, the title, synopsis, and category are extracted from the textual data. The title refers to the headline of the article on the website, which the video belongs to, while the synopsis is a longer summary of the video's content. The category defines a broader topic, such as sports or entertainment.

5 Results & Discussion

This section describes the results for the multilingual evaluation as well as for different pipelines and the human evaluation.

5.1 Multilingual Model Evaluation

Figure 2 shows the recall for English and for Dutch data. The results for all models, as well as performance per dataset, are in Appendix C.

There are 8 models with Recall@1 > 0.7 for English, but only 3 for Dutch. It can also be observed that all three models that achieve high results on Dutch data also perform well on English data. All the best-performing models on Dutch come from the OpenCLIP framework. The standard CLIP model, as well as other vision-language implementations, are outperformed by OpenCLIP, especially in the non-English domain. The model that achieves the highest performance in the OpenCLIP evaluation is beaten, even if not significantly, by the "xlm-roberta-large-ViT-H-14" model, which only scores 6th in the overall OpenCLIP evaluation. Also, none of the Multilingual CLIP models score above the OpenCLIP models, even though they are specifically trained to handle data in different languages. In the end, it was decided to use the "xlm-roberta-large-ViT-H-14" model (laion, 2024a) from OpenCLIP as it yields the best performance on Dutch data. This model uses a "ViT-H-14" (laion, 2024b) model as the image encoder and the "xlm-

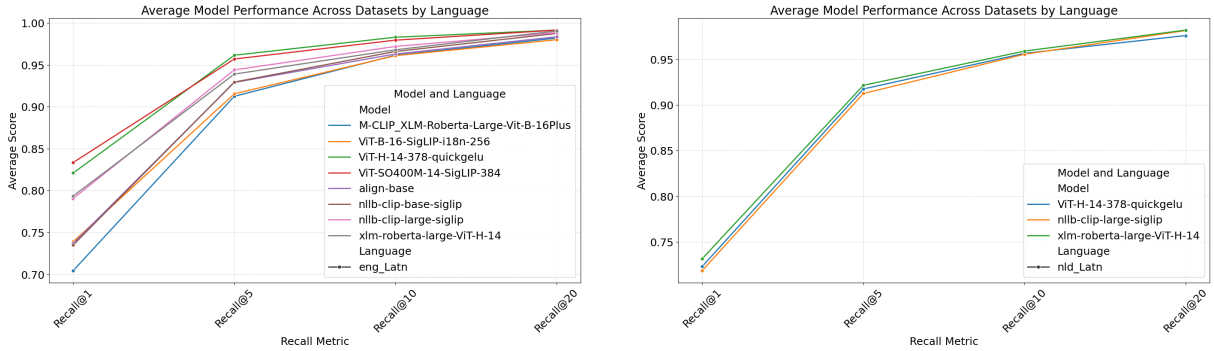

Figure 2: Left: Recall Graph for English for Top 1, 5, 10, and 20. Right: Recall Graph for Dutch for Top 1, 5, 10, and 20. Only models with Recall@1 > 0.7 are shown.

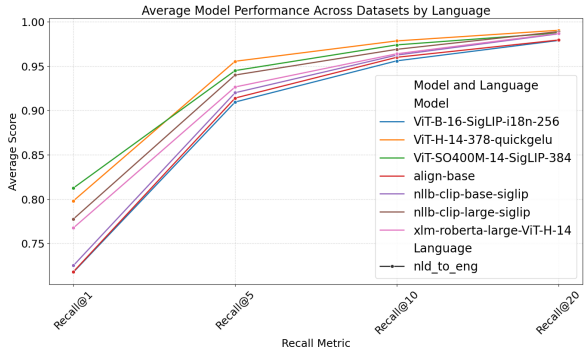

Figure 3: Recall Graph for Dutch Captions translated to English for Top 1, 5, 10, and 20. Only models with Recall@1> 0.7 are shown.

roberta-large" model (Conneau et al., 2019), which is trained on 100 languages, as the text encoder. All the following experiments are performed with this model.

The ViT-SO400M-14-SigLIP-384 model (Zhai et al., 2023) (Beyer et al., 2022) yields the best results on the translated data, as can be seen in Figure 3. Therefore, it was decided to use this model for experiments on the translated text inputs. After evaluating the Top 5 frames, it was decided not to pursue this approach further because the Top 5 frames did not differ from the non-translated approach, and it would require the extra step of translating every title.

5.2 Pipelines

In comparing the pipelines, differences in output with respect to the text input can be observed. For example, as shown in Figures 4 and 5, the model's output differs when given only the title as the text input versus when given both the title and the synopsis. Since the title of this video does not include cars, the system in Figure 4 does not rank the frames with a car in the first three places. In the synopsis, cars are mentioned, which leads the system in Figure 5 to pick these frames for a higher position. Therefore, evaluating which text input delivers the best results is an important part of this research. As described in Section 3.2, different models with different inputs have been tested. To obtain a proper comparison, A/B testing with human evaluators, as described in Section 3.3.1, must be conducted. But these preliminary results already show that the kind of text given to the model has a significant influence on which thumbnails are selected and how they are ranked.

5.3 CLIP Pipeline Testing

In the A/B test, 70 videos were evaluated by 9 human evaluators, yielding 549 labels in total, as not all evaluators finished their evaluations. Figure 6 shows how often the different models have been picked. As already mentioned in 3.3.1, when the same top frame is picked by more than one model, all of the models with that frame count as being picked. Therefore, the overall count exceeds the number of labels. As can be seen, two of the models with context, namely the one that also gives the synopsis as the text input (ID L) and the one that uses the summarized title (ID M), perform significantly better than the current system. While the current system is picked 120 times, the synopsis model is picked 219 times, and the summary model is picked 211 times. For the system that filters the frames according to their aesthetic and technical scores, the performance, with 147 times picked, is not as good as for the other two, while the system that incorporates the similarity score into the current formula performs worse than the current system, being picked only 112 times. Therefore, it seems that removing all frames with subtitles is a

Figure 4: Top 5 Frames for only Title as Text Input

Figure 5: Top 5 Frames for Title and Synopsis as Text Input

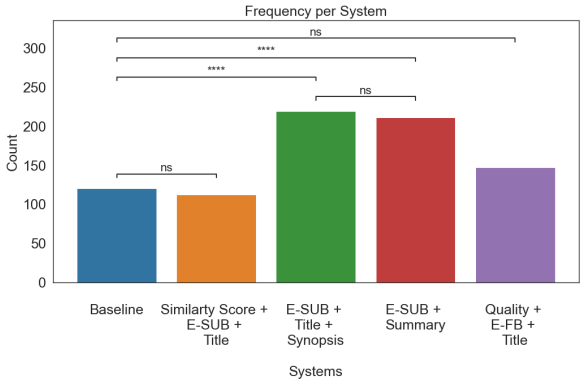

Figure 6: Frequency of how often a System is picked in the CLIP Pipeline evaluation. The lines above the bars show if the systems are significantly different. **ns** indicates $0.05 < p \leq 1.00$, * indicates $0.01 < p \leq 0.05$, ** indicates $0.001 < p \leq 0.01$, *** indicates $0.0001 < p \leq 0.001$ and **** indicates $p \leq 0.0001$.

good enough filtering step for the frames.

To determine whether the differences between the systems are statistically significant, a χ^2 Goodness-of-Fit Test is performed. This test checks whether the observed system frequencies follow a uniform distribution, indicating that any variation is due to chance rather than a meaningful difference. This test gives a p-value of 8×10^{-13}, which, for an alpha of 0.05, means that the distribution differs significantly from a uniform distribution. To see which systems actually differ significantly from each other, a χ^2 Independence Test is also performed pairwise, using Bonferroni as the p-value correction. For an alpha of 0.05, the results are shown in Figure 6. Looking at these results, the synopsis and summary models differ significantly from the other three models.

Figure 18 shows how often each of the 8 evaluators picked which system(s). Again, if two or more systems pick the same frame, the count for all systems with that frame is increased. As expected, almost all evaluators prefer the synopsis and summary frames over the rest. While the synopsis count across all evaluators is pretty similar, the summary count differs more between evaluators. Therefore, it seems that the synopsis model is the most consistent and is liked by multiple evaluators.

Figure 19 shows how often each of the 8 evaluators picked which position the image was at in the experiment. This means that if the image was the first, second, etc., in the list, then it would appear as the first, second, etc., in the list when all five thumbnails are displayed next to each other. Interestingly, it can be seen that, for almost all evaluators, the image at position 1 was selected most often. This could be because if they think that all thumbnails are equally fitting, they tend to just pick the first one. Especially evaluator 4 has a big preference for the first image. Even after excluding these results, there is no significant difference between the systems. Because the image order is shuffled for every experience, even though there is a preference for image 1, the results are still feasible, since the same system is not always in the first position.

To measure agreement among evaluators, Krippendorff's Alpha is calculated. The result is roughly 0.4, indicating that, at least some of the time, no evaluator agrees with the others. This again shows that picking a thumbnail is a highly subjective task, and that evaluators' opinions can differ widely about what a good thumbnail should look like.

5.4 Fine-Tuning

On the test set, the Recall@1 improves from 24.9 to 30.9 for the finetuned model. Furthermore, the Recall@5 improves from 43.9 to 55.1, and the Recall@10 increases from 52 to 64.9. This shows

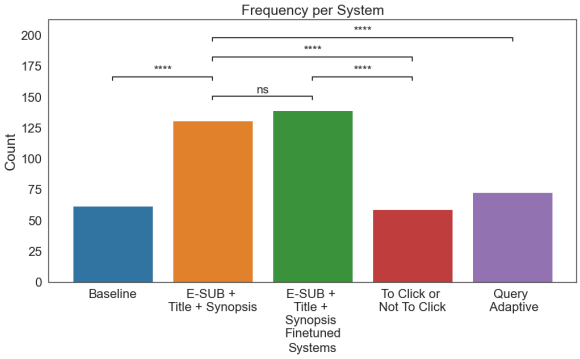

Figure 7: Frequency of how often a System is picked in the expert evaluation. The lines above the bars show if the systems are significantly different. **ns** indicates 0.05 < p ≤ 1.00, * indicates 0.01 < p ≤ 0.05, ** indicates 0.001 < p ≤ 0.01, *** indicates 0.0001 < p ≤ 0.001 and **** indicates p ≤ 0.0001.

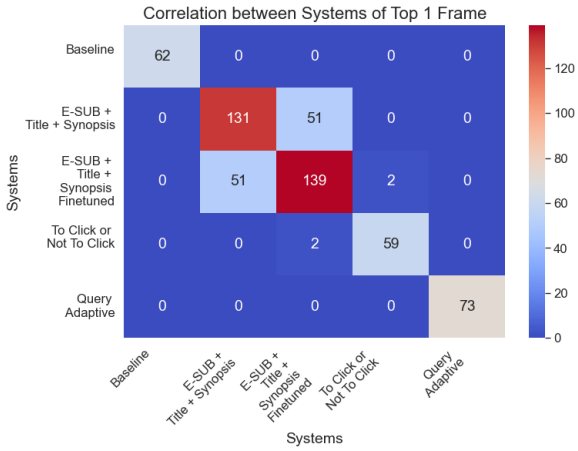

Figure 8: Top 1 Frame Correlation between different Systems in the human expert evaluation. Blue indicates low correlation, white medium and red high correlation.

that fine-tuning the model improves matching on domain-specific data and identifies the correct image more often than the non-fine-tuned model.

5.5 Expert Testing

To ensure the videos used in the CLIP Pipeline evaluation are not biased toward the CLIP system, a new dataset of 100 news and entertainment videos was used for the expert evaluation. In total, 411 evaluations could be acquired from the experts, as not everyone finished the evaluation. Including the duplications of the Top 1 frame a total of 464 labels was acquired. The evaluation setup was the same as described in Section 5.3, taking into account whether two models produced the same Top 1 frame, as well as the setup for the statistical testing.

Figure 7 shows the results of the expert evaluation. Interestingly, the CLIP and its fine-tuned version perform very similarly across the dataset with 139 respectively 131 times picked. It can also be seen that the baseline model, as well as both reference implementations, perform significantly worse than both of our CLIP models with 73 picks for the query adaptive implementation, 62 for the current model and 59 for the To Click or no to click baseline, highlighting the usefulness of CLIP for thumbnail selection. Also, as the performance of the non-finetuned model does not differ significantly from that of the fine-tuned one, data- and resource-intensive fine-tuning is not necessary to produce satisfying results.

The high correlation between the two CLIP Systems could be explained when looking at the system correlation in Figure 8. 51 times, both systems picked the same frame, resulting in both systems being chosen.

Out of the 411 evaluations the question if the picked frame is an acceptable thumbnails was answered with *yes* 337 times. When excluding the ones answered with *no* the results in terms of significant difference do not change.

6 Conclusion

According to human experts comparing different automatic tools for picking a suitable thumbnail for a video, CLIP proves to be a useful approach, outperforming the current as well as two baseline systems significantly.

An important part in getting satisfying results is the capability of the model to be able to process Dutch text which can be a challenge for models not trained, or not trained properly, on non-English texts.

Furthermore, it shows that having a proper frame filtering approach in place increases the performance of the CLIP models especially with their text spotting bias and therefore being drawn to images with text to them which is not desirable for thumbnails. Therefore, our approach of only checking for subtitles and not text in general on frames for filtering proved to be very effective.

All in all, including textual context into automatic thumbnail selection can provide satisfying results even for non-English languages. Future work could include better frame filtering or include an aesthetic analysis directly into the CLIP model. Lastly, evaluation can be extended to include languages other than Dutch.

Limitations

Thumbnail selection is a highly subjective task. Therefore, multiple evaluators have been asked to evaluate the data to get a variety of opinions. But even with multiple evaluators per video, this task remains hard to evaluate, as setting a gold standard thumbnail is often also not feasible as it would again be picked due to the subjective preferences of the people creating the dataset.

Furthermore, all CLIP pipelines used frames filtered at fixed intervals as input, and no further experiments were conducted with different sampling strategies. We acknowledge that the frame-sampling strategy can significantly affect the selected frame. However, due to data availability, we had to stick to this strategy.

The non-expert evaluators mentioned that, in some cases, they did not like any of the shown thumbnails. This could be either due to the expectations of how a thumbnail should look, for example, it should be custom-made to grab attention, like is common on YouTube, or due to all of the systems outputting non-fitting images. As there was no option for saying "no thumbnail was good", it may be that, in that case, the evaluators just picked the first image. Therefore, in the expert evaluation, we included the option to say if the thumbnail they picked is fitting or not, but did not make another run with non-experts to evaluate the CLIP Pipeline with this option.

Also, the scope of this work was limited to Dutch data. While the performance for the same task in English should intuitively be higher, no further experiments for languages other than Dutch have been conducted.

Lastly, all experiments were conducted on data from one Dutch media company. Therefore, further work can include testing the approach on different kinds of videos, e.g., YouTube or other news sites, to see how it impacts performance.

References

Evlampios Apostolidis, Georgios Balaouras, Vasileios Mezaris, and Ioannis Patras. 2023. Selecting a diverse set of aesthetically-pleasing and representative video thumbnails using reinforcement learning. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 2460–2464.

Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. 2022. Big vision. https://github.com/google-research/big_vision.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Elia Fantini. 2024. Automating video thumbnails selection and generation with multimodal and multistage analysis. *Preprint*, arXiv:2410.19825.

Yuli Gao, Tong Zhang, and Jun Xiao. 2009. Thematic video thumbnail selection. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 4333–4336.

Samuel Hoffstaetter. 2024. pytesseract. https://pypi.org/project/pytesseract/, accessed on 31-10-2024.

Andreas Husa, Cise Midoglu, Malek Hammou, Steven A. Hicks, Dag Johansen, Tomas Kupka, Michael A. Riegler, and Pål Halvorsen. 2022. Automatic thumbnail selection for soccer videos using machine learning. In *Proceedings of the 13th ACM Multimedia Systems Conference*, MMSys '22, page 73–85, New York, NY, USA. Association for Computing Machinery.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *Preprint*, arXiv:2102.05918.

Rakpong Kittinaradorn. 2024. easyocr. https://pypi.org/project/easyocr/, accessed on 31-10-2024.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Preprint*, arXiv:2309.04662.

laion. 2024a. Clip vit-h/14 frozen xlm roberta large - laion-5b. https://huggingface.co/laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k, accessed on 31-10-2024.

laion. 2024b. laion/clip-vit-h-14-laion2b-s32b-b79k. https://huggingface.co/laion/CLIP-ViT-H-14-laion2B-s32B-b79K, accessed on 31-10-2024.

LAION-AI. 2024. Clip benchmark. https://github.com/LAION-AI/CLIP_benchmark, accessed on 31-10-2024.

Yiqi Lin, Conghui He, Alex Jinpeng Wang, Bin Wang, Weijia Li, and Mike Zheng Shou. 2024. Parrot captions teach clip to spot text. *Preprint*, arXiv:2312.14232.

Wu Liu, Tao Mei, Yongdong Zhang, Cherry Che, and Jiebo Luo. 2015. Multi-task deep visual-semantic embedding for video thumbnail selection. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3707–3715.

Joanna Materzyńska, Antonio Torralba, and David Bau. 2022. Disentangling visual and written concepts in clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16410–16419.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Yale Song, Miriam Redi, Jordi Vallmitjana, and Alejandro Jaimes. 2016. To click or not to click: Automatic selection of beautiful thumbnails from videos. *Preprint*, arXiv:1609.01388.

Arun Balajee Vasudevan, Michael Gygli, Anna Volokitin, and Luc Van Gool. 2017. Query-adaptive video summarization via quality-aware relevance estimation. *CoRR*, abs/1705.00581.

YouTube. Youtube official blog (press).

Zhifeng Yu and Nanchun Shi. 2021. A multi-modal deep learning model for video thumbnail selection. *CoRR*, abs/2101.00073.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. *Preprint*, arXiv:2111.07991.

A Expert Testing

Figure 9 shows how often each system was shown at each position after shuffling for each experiment to ensure randomness in their placement.

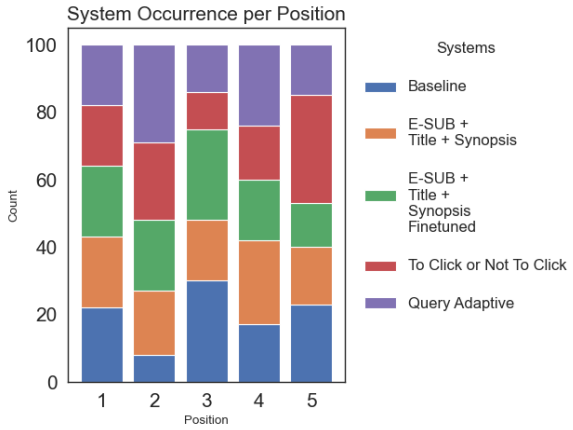

Figure 9: Shuffling Results. How often each system appears at which position. For every video the order of frames shown to the evaluators was changed to mitigate the position bias.

Figure 10 shows how often each evaluator picked which system.

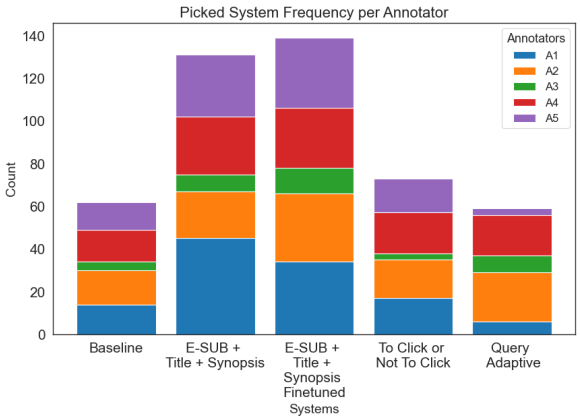

Figure 10: Individual Analysis per Annotator. Shows how often each annotator picked each system.

Figure 11 shows how often each position was picked by each evaluator.

B Baseline Proprietary System

The current system uses image aesthetics, technical image scores, text detection, face detection, and face clustering. It ranks the frames by their accumulated scores across all these categories. For image aesthetics and technical quality, MobileNet is used, which assigns a score to each frame in each category. For text detection, the score is inversely

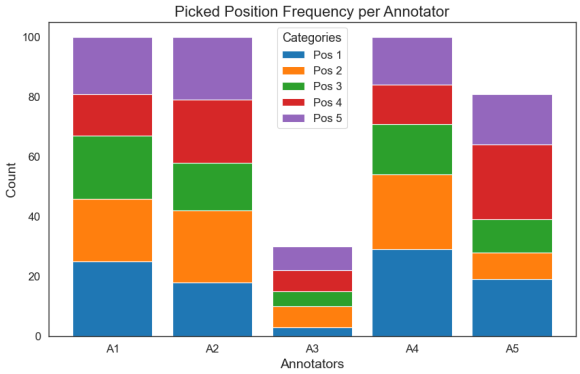

Figure 11: Individual Position Analysis per Annotator. Shows how often each annotator picked each position.

proportional to the length of the text, meaning the more words there are in the image, the lower the score gets. Also, the score is boosted by 1.2 if the image contains faces. This leads the current system to nearly always pick frames in which a person or a group of people is being interviewed or is otherwise in front of the camera. Especially the face boost and face clustering boost have a huge impact on the final score. Having faces on the thumbnail may be beneficial if the video is about a famous person, but not if it is, for example, about a gallery opening and the people interviewed are random visitors. In this case, showing an interviewee on the thumbnail may indicate nothing about the video content, but an image of the gallery would tell us that this video has something to do with a gallery (opening).

C Multilingual Model Evaluation

Figure 12 shows the benchmarking results for all the evaluated models. Table 3 shows the results for all models in Dutch.

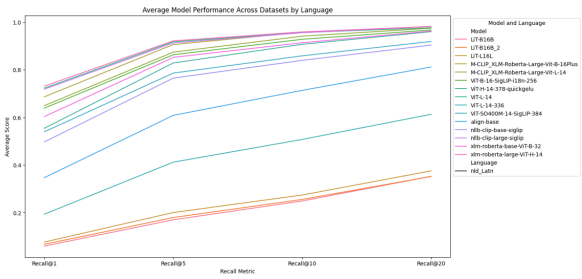

Figure 12: Recall Graph for Dutch for all Models For Top 1, 5, 10, and 20

Figure 13 shows a comparison of each model for each language and dataset.

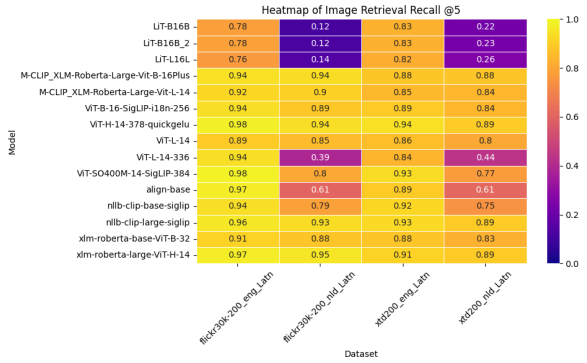

Figure 13: Performance for Recall@5 for every model on each dataset and language. Blue indicates low performance, red indicates mid performance and yellow indicates high performance for the model/dataset/language combination.

D OCR

It is important that the thumbnails do not have (a lot of) text on them. Text can occur in the form of subtitles, graphics (like name inserts), or logos, for example, on a t-shirt a person being interviewed is wearing. While evaluating the frames without text as detected by the text-detection system already in place, it was observed that many texts (especially burned-in subtitles) were not detected. The current system utilizes "pytesseract" (Hoffstaetter, 2024) to detect text on images. Although it is often recommended, it does not appear to function well on the video frames. Therefore, it was decided to implement another text detection system to improve detection performance. In the end, "easyocr" (Kittinaradorn, 2024) was used because it detected text on frames more often than "pytesseract".

E Technical & Aesthetic Scores

Figures 14 and 15 show the distribution of the technical and aesthetic scores between the Top 5 and non-Top 5 frames. The vertical line shows the chosen cutoff value.

F CLIP Pipeline Testing

The following shows the original Dutch prompt as well as its English translation given to the human evaluators for evaluating the CLIP Pipelines.

Original Prompt (Dutch)

Uit de bijlagen, selecteer de afbeelding die je de beste thumbnail voor een video met de titel {titel} vindt. Het gaat om een video die op <media website> staat als een nieuws of entertainment artikel.

Model	R@1	R@5	R@10	R@20
LiT-B16B	0.0595	0.1700	0.2490	0.3525
LiT-B16B_2	0.0670	0.1795	0.2555	0.3520
LiT-L16L	0.0765	0.2005	0.2740	0.3755
M-CLIP_XLM-Roberta-Large-Vit-B-16Plus	0.6865	0.9060	0.9575	0.9810
M-CLIP_XLM-Roberta-Large-Vit-L-14	0.6495	0.8740	0.9410	0.9745
ViT-B-16-SigLIP-i18n-256	0.6390	0.8630	0.9280	0.9670
ViT-H-14-378-quickgelu	0.7230	0.9175	0.9565	0.9760
ViT-L-14	0.5545	0.8285	0.9070	0.9600
ViT-L-14-336	0.1935	0.4120	0.5075	0.6130
ViT-SO400M-14-SigLIP-384	0.5395	0.7860	0.8585	0.9190
align-base	0.3470	0.6085	0.7135	0.8115
nllb-clip-base-siglip	0.4970	0.7650	0.8395	0.9035
nllb-clip-large-siglip	0.7185	0.9125	0.9555	0.9815
xlm-roberta-base-ViT-B-32	0.6035	0.8520	0.9135	0.9625
xlm-roberta-large-ViT-H-14	**0.7315**	**0.9215**	**0.9590**	**0.9820**

Table 3: Image–text retrieval performance for Dutch. R@k denotes Recall@k.

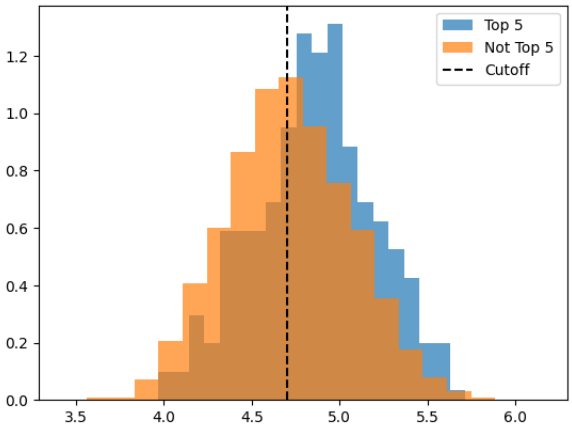

Figure 14: **Blue** indicates the distribution of the technical scores of the Top 5 frames according to the baseline. **Orange** indicates the distribution of the non-Top 5 frames according to the baseline. The vertical line shows the cutoff point. Frames that scored above this threshold were included when filtered by quality, frames that scored below were not included.

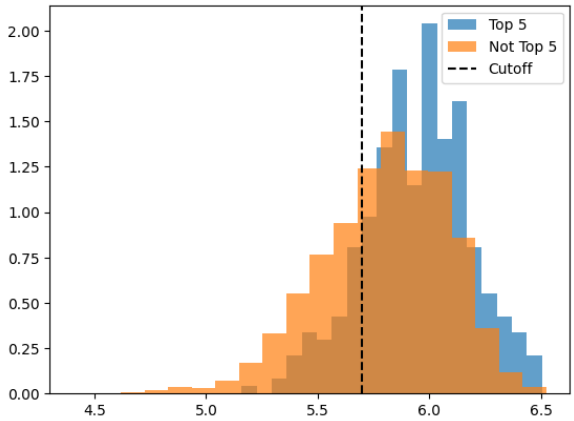

Figure 15: **Blue** indicates the distribution of the aesthetics scores of the Top 5 frames according to the baseline. **Orange** indicates the distribution of the non-Top 5 frames according to the baseline. The vertical line shows the cutoff point. Frames that scored above this threshold were included when filtered by quality, frames that scored below were not included.

English Translation

From the attachments, select the image that you think is the best thumbnail for a video with the title {title}. The video will be shown on <media website> as a news or entertainment article.

Figure 16 shows the correlation between different models for how often, when an evaluator picks a frame, another model has the same Top 1 frame as the picked one. As it can be seen, there is a high correlation between the synopsis and summary model, meaning they often pick the same Top 1 frame, while for all the other models, the number of correlations is way lower.

Figure 17 shows how often each system was shown at each position after shuffling for each experiment to ensure randomness in their placement.

Figure 18 shows how often each evaluator picked which system.

Figure 19 shows how often each position was picked by each evaluator.

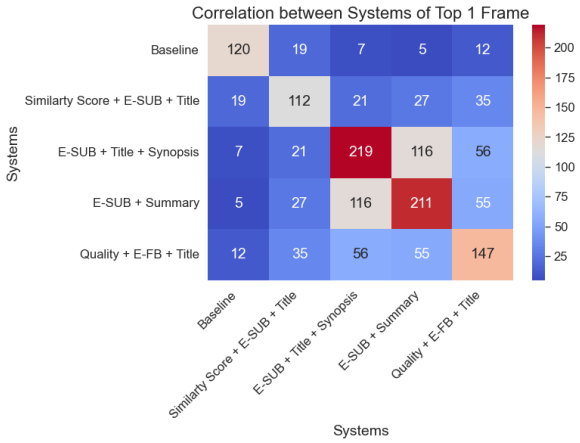

Figure 16: Top 1 Frame Correlation between different Systems in the CLIP Pipeline evaluation. Blue indicates low correlation, white medium and red high correlation.

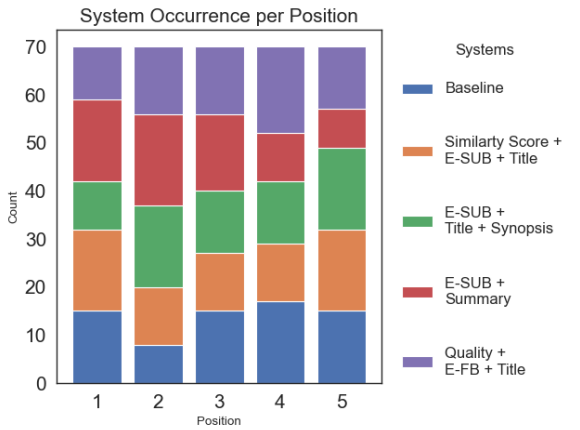

Figure 17: Shuffling Results. How often each system appears at which position. For every video the order of frames shown to the evaluators was changed to mitigate the position bias.

G Human Evaluators

The evaluators for the CLIP pipeline were annotators in a transitional job program contracted through a third party for a media company. This third party works with the local municipality and provides work coaching for people who face barriers to enter the labour market. The program is intended to acquire work experience, working as annotators in this transitional role, with the ultimate goal of preparing them for their transition to the job market, under the careful supervision of job coaches. Note that all the annotators fall under the responsibility of the UWV and retain their benefits while participating in the aforementioned program.

The expert evaluators are employed by a media company and did the evaluation during paid working hours.

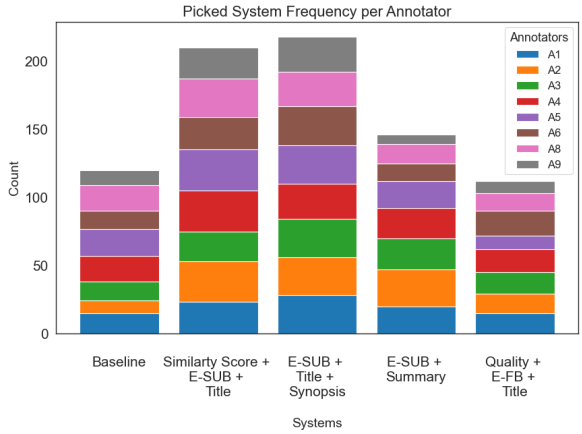

Figure 18: Individual Analysis per Annotator. Shows how often each annotator picked each system.

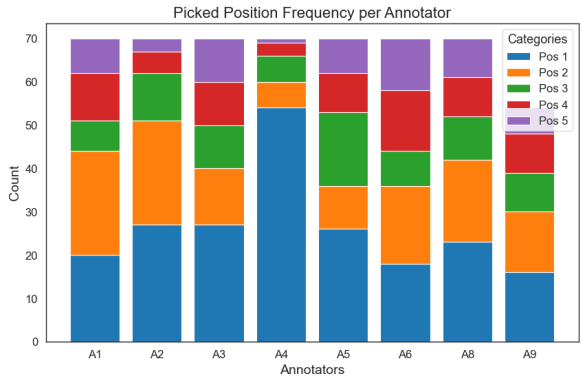

Figure 19: Individual Position Analysis per Annotator. Shows how often each annotator picked each position.Frequency of what position which evaluator prefers