ICIP 2017

Progressive Communication for Interactive Light Field Image Data Streaming

Eduardo Peixoto, Bruno Macchiavello, Edson Mintsu Hung, Camilo Dorea and Gene Cheung

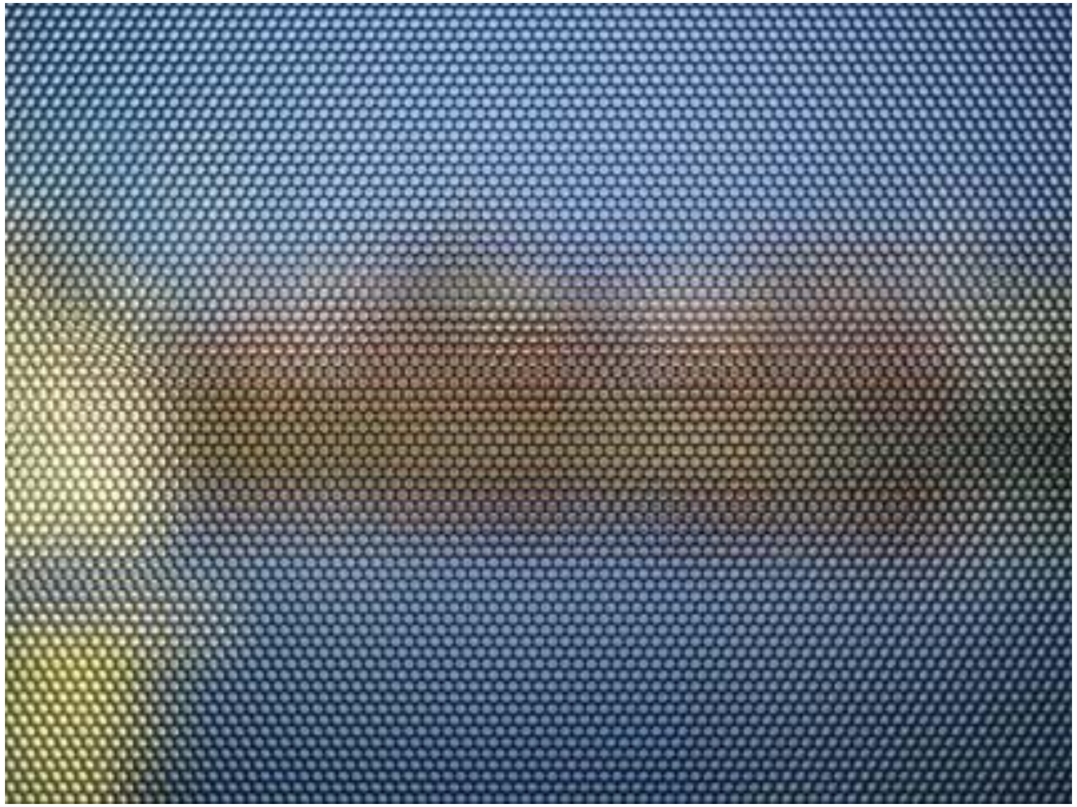
fapdf
Fundação de Apoio à Pesquisa do Distrito Federal

CNPq

Universidade de Brasília

NII
大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

Introduction



Raw Light-Field Image

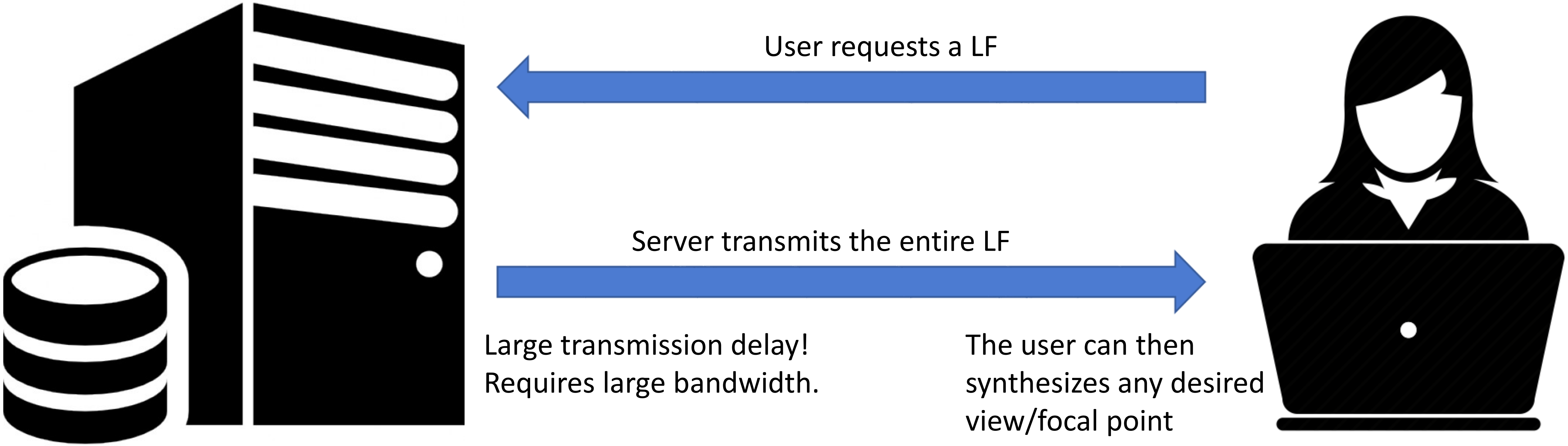
http://cameramaker.se/plenoptic.htm

Changing the focal point

Changing the view point

How to transmit Light-Field data?

1 – Transmits the entire Light-field data.



User requests a LF

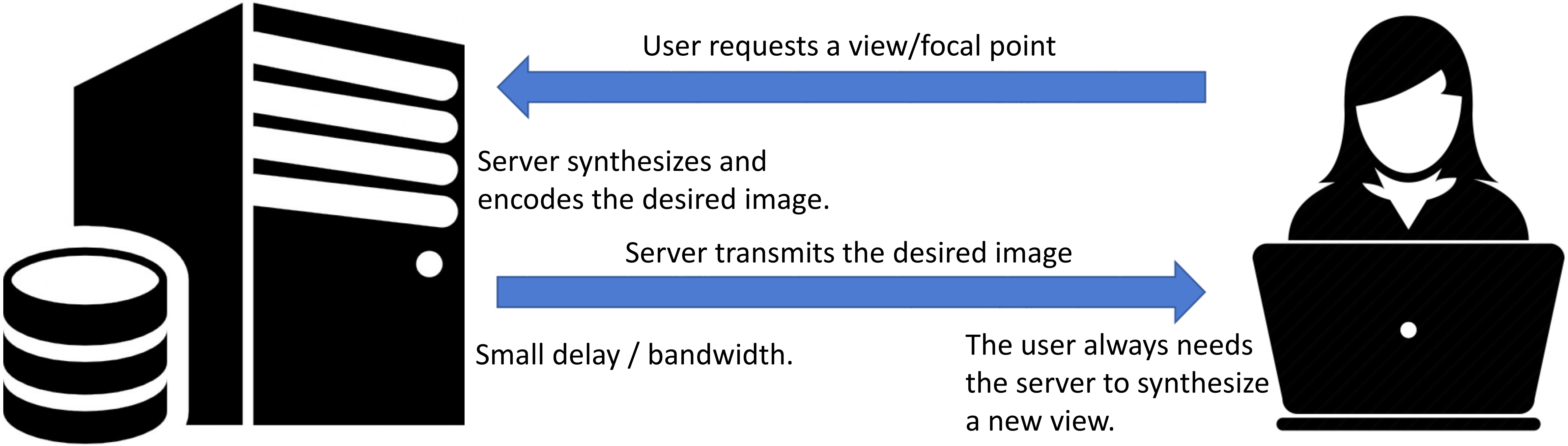
Server transmits the entire LF

Large transmission delay!
Requires large bandwidth.

The user can then
synthesizes any desired
view/focal point

How to transmit Light-Field data?

2 - Interactive Light Field Streaming (ILFS)



User requests a view/focal point

Server synthesizes and
encodes the desired image.

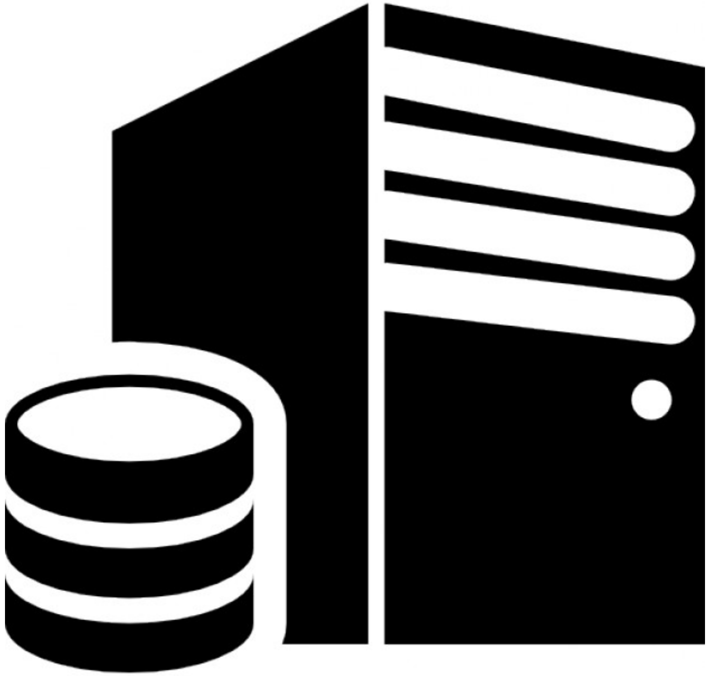
Server transmits the desired image

Small delay / bandwidth.

The user always needs
the server to synthesize
a new view.

How to transmit Light-Field data?

The proposed work: **Progressive Light-Field Communication (PLFC)**

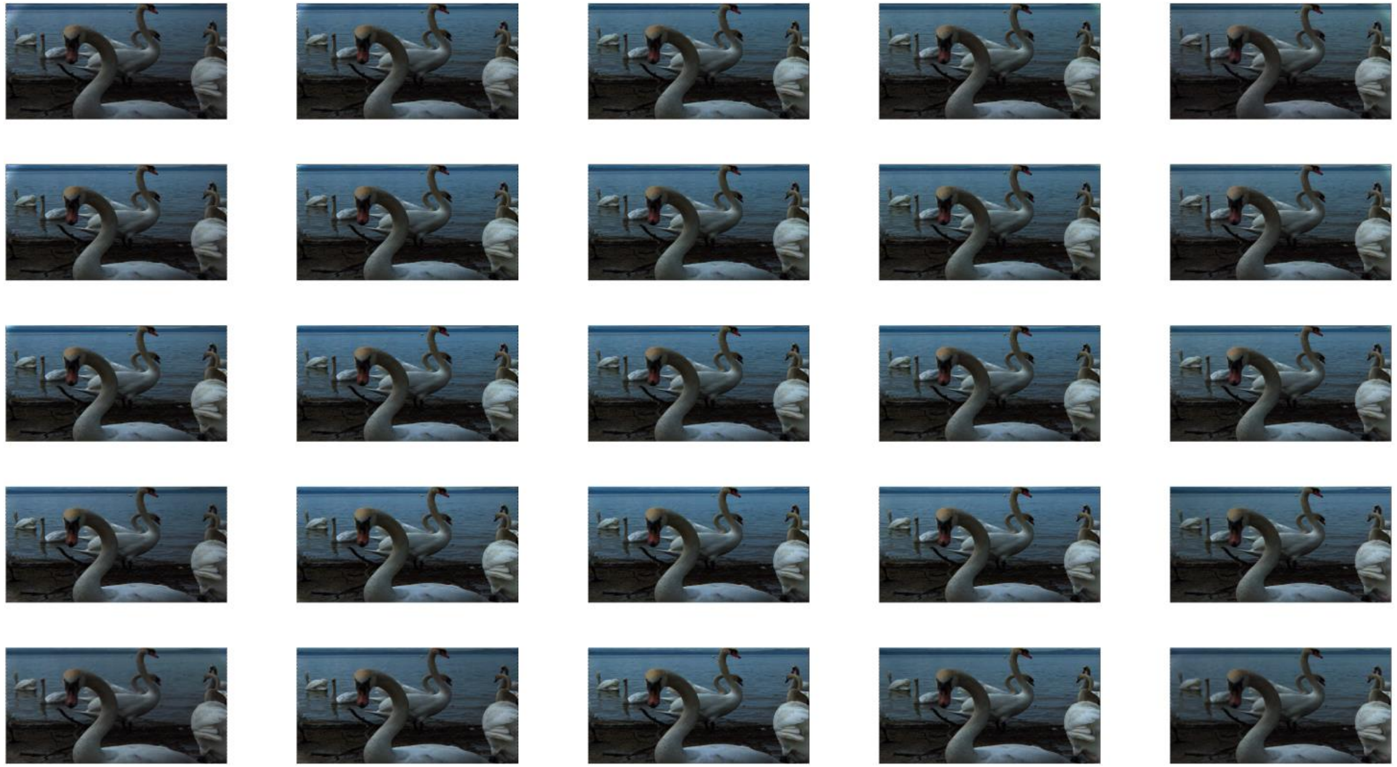


In the ILFS approach, the user never learns anything about the LF.

In our approach, for each received synthesized image, the client “decodes” and recovers a new sub-aperture image using a cache of known sub-aperture images.

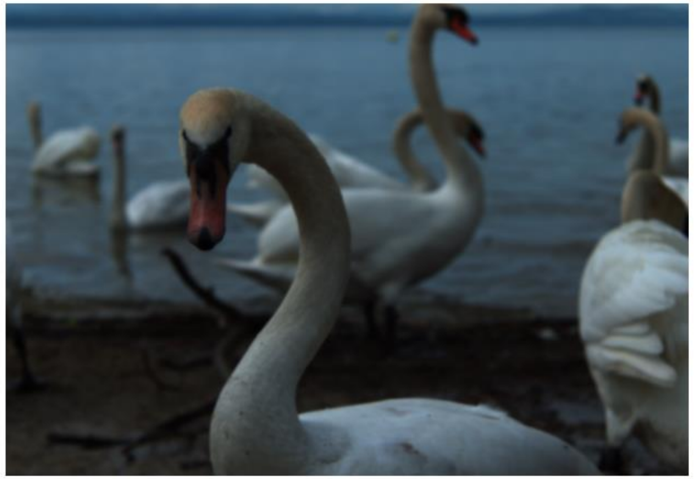
It fits between the two previous approaches.

A quick overview of LF Synthesis



Sample Sub-Aperture Images (of a total of 225)

$$\mathbf{v}(\alpha, \mathcal{S}^0) = \frac{\sum_{i \in \mathcal{S}^0} w_i^{\alpha} \mathbf{x}_i^{\alpha}}{\sum_{i \in \mathcal{S}^0} w_i^{\alpha}}$$



Synthesized view

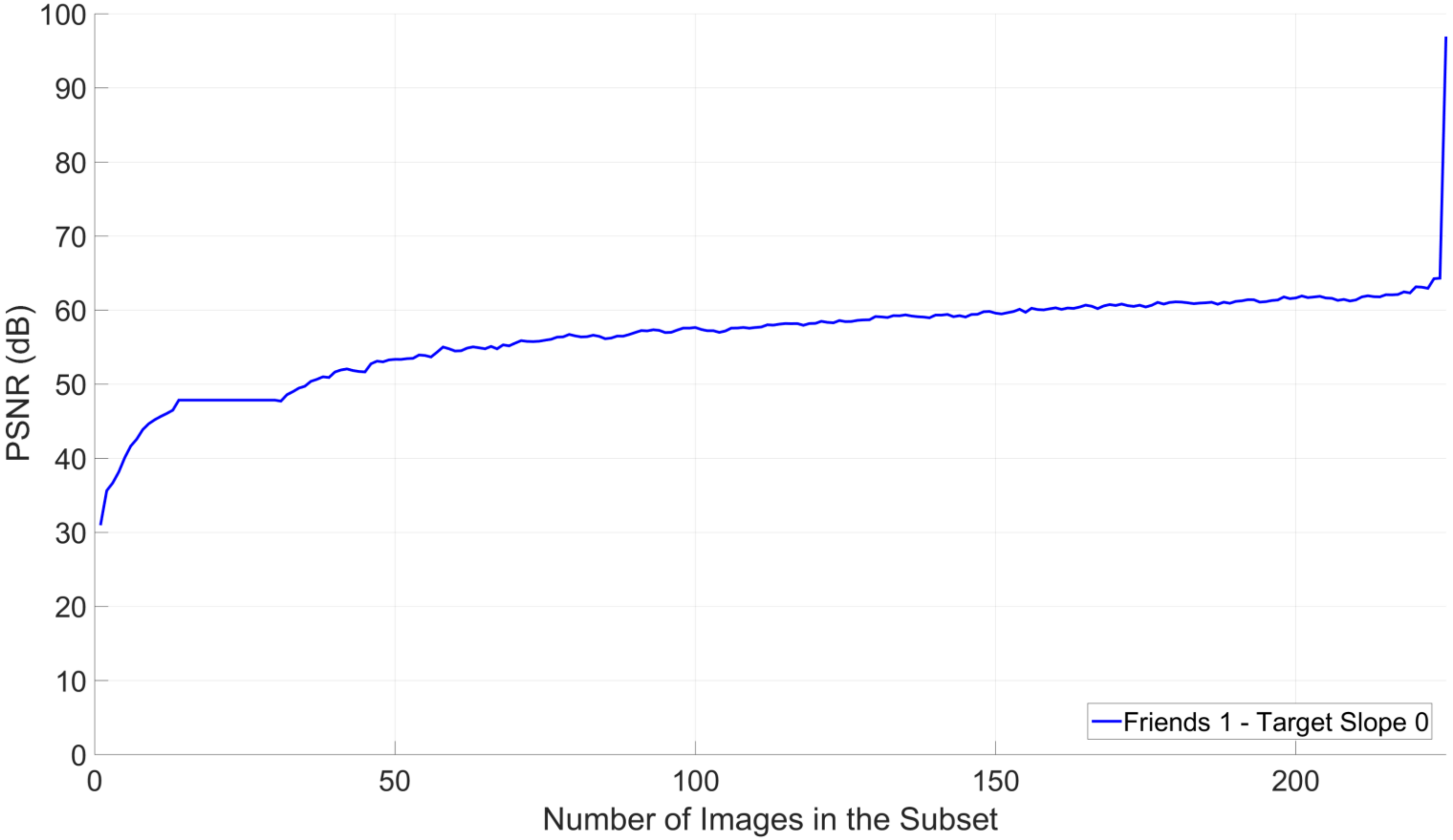
A quick overview of LF Synthesis

- What happens if we use a subset $\mathcal{S}$ of the complete set $\mathcal{S}^0$ to synthesize new images?
- We can allow scalar coefficients in the linear combination.

$$\mathbf{v}(\alpha, \mathcal{S}\ ) = \frac{\sum_{i\in\mathcal{S}} p_i w_i^{\alpha} \mathbf{x}_i^{\alpha}}{\sum_{i\in\mathcal{S}} w_i^{\alpha}}$$

- How many images are needed to get a good quality?

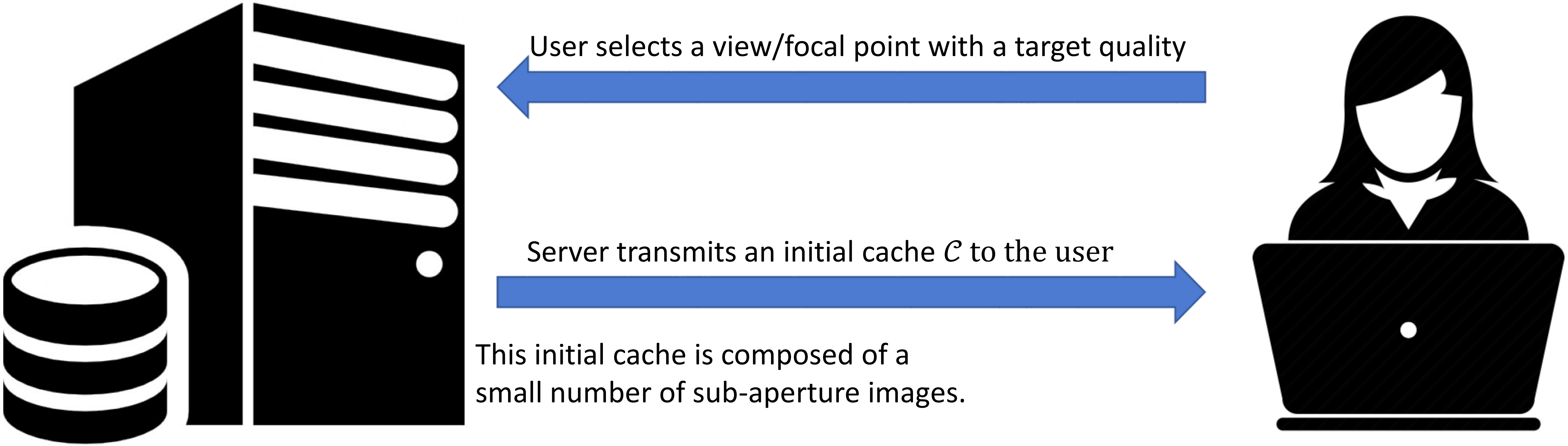
A quick overview of LF Synthesis



PSNR	Number
35 dB	2
40 dB	5
45 dB	10
50 dB	36

Interactive Transmission Framework

Initialization



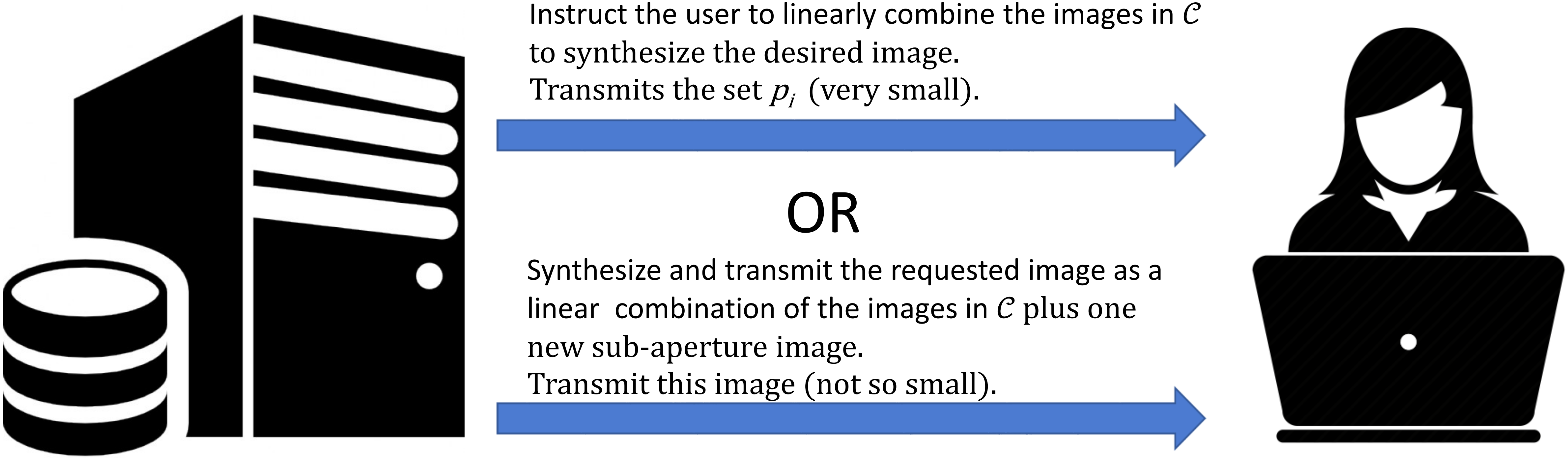
User selects a view/focal point with a target quality

Server transmits an initial cache $\mathcal{C}$ to the user

This initial cache is composed of a
small number of sub-aperture images.

Interactive Transmission Framework

The encoder then has two options:



Instruct the user to linearly combine the images in $\mathcal{C}$ to synthesize the desired image.
Transmits the set p_i (very small).

OR

Synthesize and transmit the requested image as a linear combination of the images in $\mathcal{C}$ plus one new sub-aperture image.
Transmit this image (not so small).

Interactive Transmission Framework

- How to choose between these options?
  - If the encoder is able to meet the user's defined quality using only the images in $\mathcal{C}$, then this option is used.
  - Otherwise, we use the "Synthesize and transmit" iteratively.

Synthesize and Transmit

- When this option is used, the encoder synthesizes a view using the images in the cache $\mathcal{C}$ plus one new image $\mathbf{z}$.

$$\mathbf{v}(\alpha, \mathcal{C} \cup \mathbf{z}\,) = \frac{\sum_{i\in\mathcal{C}} (p_i w_i^{\alpha} \mathbf{x}_i^{\alpha}) + p_z w_z^{\alpha} \mathbf{z}^{\alpha}}{\sum_{i\in\mathcal{C}} w_i^{\alpha} + \, w_z^{\alpha}}$$

- Once the decoder receives $\mathbf{v}(\alpha, \mathcal{C} \cup \mathbf{z}\,)$, it can estimate the new image $\mathbf{z}$.
- This image is then added to the cache, so it “learns” something new about the Light-Field.

Synthesize and Transmit

- We actually transmit the residual $\mathbf{v}(\alpha, \mathcal{C} \cup \mathbf{z}\,)$ - $\mathbf{v}(\alpha, \mathcal{C}\,)$

$$\mathbf{v}(\alpha, \mathcal{C} \cup \mathbf{z}\,) = \frac{\sum_{i \in \mathcal{C}} (p_i w_i^\alpha \mathbf{x}_i^\alpha) + p_z w_z^\alpha \mathbf{z}^\alpha}{\sum_{i \in \mathcal{C}} w_i^\alpha + \; w_z^\alpha}$$

$$\mathbf{v}(\alpha, \mathcal{C}\,) = \frac{\sum_{i \in \mathcal{C}} q_i w_i^\alpha \mathbf{x}_i^\alpha}{\sum_{i \in \mathcal{C}} w_i^\alpha}$$

- We transmit this residual using H.264/AVC with zero MVs.
- Since the decoder actually receives $\mathbf{v}'(\alpha, \mathcal{C} \cup \mathbf{z}\,)$, it can only estimate $\mathbf{z}$. Thus, to avoid drift, all of these actions are mimicked by the encoder.

Synthesize and Transmit

- Note that it is not guaranteed that we will achieve the desired quality by adding just one image to the cache.

- Thus, we use this framework in an iterative way – each time, we send one new image, which is added to the cache. If the quality is not met, we send another one, until this target quality is met.

How to initialize the Cache?

- Depending on the desired quality, we set a number of sub-aperture images in the initial cache.
- We then use a greedy algorithm, in which we add to the cache the image that is the most benefitial until this budget is met.

Target PSNR	Size of the Initial Cache
36 dB	3
38 dB	5
40 dB	8

- The cache is transmitted using H.264/AVC with an IPP structure.

How to choose the new image $\mathbf{z}$?

- The key idea is to choose a sub-aperture image that is benefitial not only to the current requested view/focal point, but also to the expected future view/focal points.

- We have modelled the user’s interaction using a probability transition matrix. In a real scenario, the server can gather data in order to optimize this matrix.

How to choose the new image **z**?

- We can compute how much each sub-aperture image **z** improves the current view **v**.

- Thus we can define the immediate benefit:

$$B_{\mathbf{v}}^{1}(\mathcal{C},\mathbf{z}) = D_{\mathbf{v}}(\mathcal{C}) - D_{\mathbf{v}}(\mathcal{C} \cup \mathbf{z})$$

How to choose the new image **z**?

- But we also **z** to be beneficial to future view/focal points.
- We define the importance of a view *i* as : $\theta(i, \mathbf{v}, t) = \sum_{\tau=1}^{T-t} [\mathbf{1}_{\mathbf{v}} \mathbf{P}^{\tau}]_i$
- And compute this second benefit as:

$$B_{\mathbf{v}}^2(\mathcal{C}, \mathbf{z}, \mathcal{V}) = \sum_{\mathbf{u} \in \mathcal{V}^0 \setminus \mathcal{V}} \theta(\mathbf{u}, \mathbf{v}, t) \big(D_{\mathbf{u}}(\mathcal{C}) - D_{\mathbf{u}}(\mathcal{C} \cup \mathbf{z}) \big)$$

How to choose the new image **z**?

- We combine the two criteria as:

$$\max_{z \in \mathcal{S}^0 \backslash \mathcal{C}} B_{\mathbf{v}}^1(\mathcal{C}, \mathbf{z}) + \mu B_{\mathbf{v}}^2(\mathcal{C}, \mathbf{z}, \mathcal{V})$$

- Where μ is a parameter that trades off current and future considerations. For this paper, we used a small μ (μ=0.1).

Results

- We have used some images in the JPEG Pleno Dataset, taken with a Lytro Illum camera (225 sub-aperture images).

- We generated 100 focal points using a Gaussian process.

- We compare to an ILFS method using H.264/AVC in RGB mode. Subsequent images are encoded differentially (with full ME).

Results

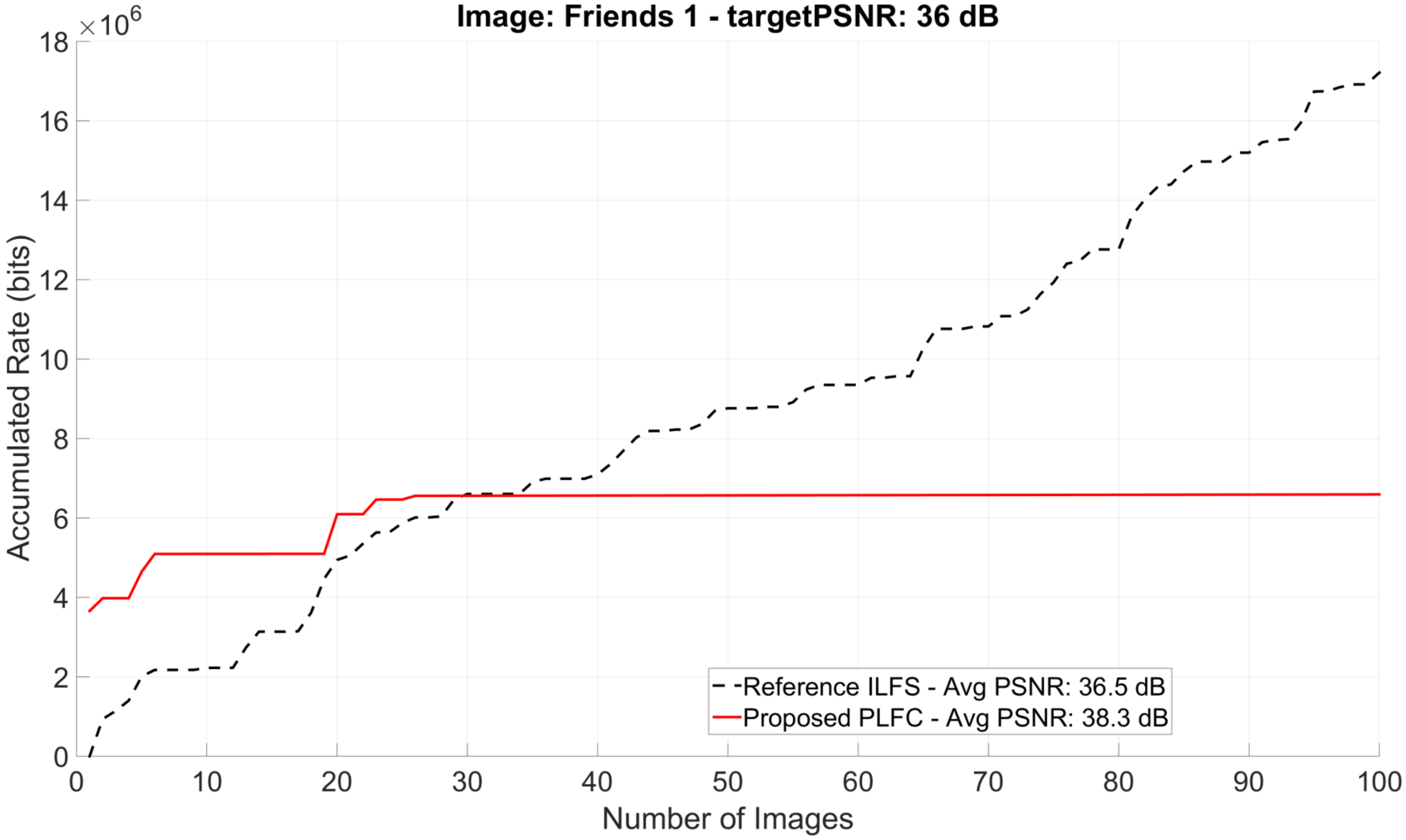
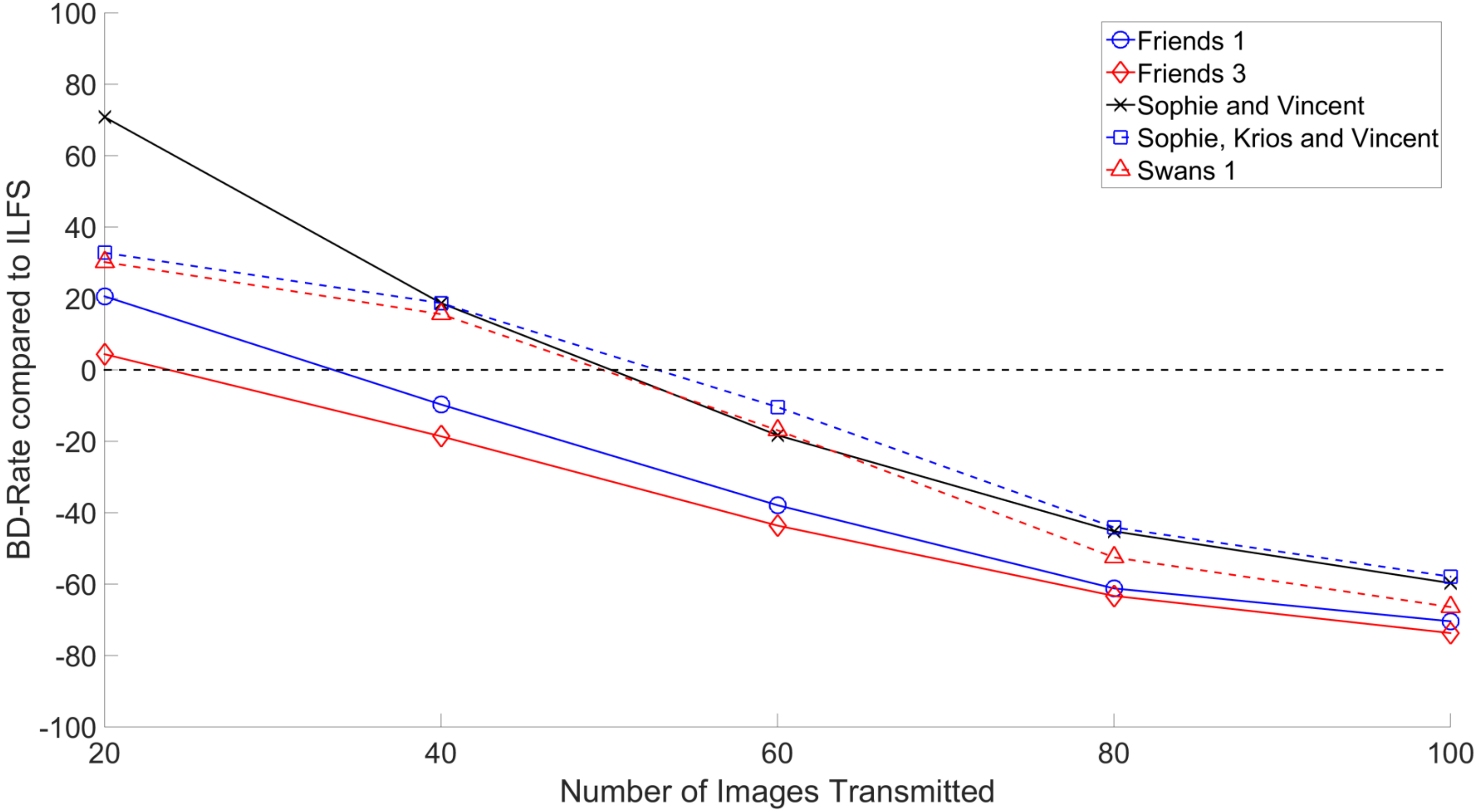


Image: Friends 1 - targetPSNR: 36 dB
×10^6
18
16
14
12
10
8
6
4
2
0
Accumulated Rate (bits)
0
10
20
30
40
50
60
70
80
90
100
Number of Images
Reference ILFS - Avg PSNR: 36.5 dB
Proposed PLFC - Avg PSNR: 38.3 dB

Results



Friends 1
Friends 3
Sophie and Vincent
Sophie, Krios and Vincent
Swans 1
BD-Rate compared to ILFS
100
80
60
40
20
0
-20
-40
-60
-80
-100
20
40
60
80
100
Number of Images Transmitted

Future Work

- There is still room for many improvements:
  - An algorithm for optimal selection of the Lagrangian multiplier that balances the immediate and future benefits.
  - A cache initialization method with variable size.
  - A more robust RD function for option selection during encoding.

ICIP 2017

Thank you

Questions?

fapdf Fundação de Apoio à Pesquisa do Distrito Federal

CNPq

Universidade de Brasília

NII 大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 National Institute of Informatics